

Morality, prejudice and cloning

The debate in the United States on human cloning took significant steps forward last week. But whether President Bush's ethics panel will serve the public well remains to be seen.

An ironic conjunction occurred last week in Washington DC, when two bodies of thoughtful people held two very different assemblies about human cloning.

The National Academies' Committee on Science, Engineering, and Public Policy presented their report, *Scientific and Medical Aspects of Human Reproductive Cloning* — the outcome of deliberations that included a near-farcical meeting attended by the maverick would-be cloner, Severino Antinori. The committee has demolished any suggestion that there is a scientific justification at present for attempting human reproductive cloning (see page 351). Their valuable review of experiments (see www.nationalacademies.org) amply demonstrates that safety issues alone are sufficient to make such research ethically unacceptable. They also clearly illustrate the differences between human reproductive cloning and the use of nuclear transplantation to produce stem cells (sometimes called therapeutic cloning).

The report also goes over some of the ground previously covered by President Bill Clinton's National Bioethics Advisory Commission (NBAC) in its 1997 report on human cloning (see <http://bioethics.georgetown.edu/nbac/pubs.html>), in examining options for discouraging human reproductive cloning. A federal ban on people's freedoms is at best a last resort. They nevertheless endorse it in this case. But, like the now-defunct NBAC, they insist on a 'sunset' provision that any such ban should be reviewed after five years.

The National Academies have refrained from delving into ethical arguments, and it is there that the science of cloning will encounter its deepest opposition. Ethical consideration is essential, but the character of ethical arguments, and how to respond to them, merits reflection. And one questionable aspect was vividly on display at the other meeting in Washington last week, the first open discussion on human cloning by President George W. Bush's Council on Bioethics.

In an earlier essay on the ethics of genetic technologies (see www.rand.org/publications/MR/MR1139/MR1139.appf.pdf), the council's chairman, Leon Kass, characterized the NBAC's recommendations on reproductive cloning as "apparently believing there are no other moral arguments [than safety] sufficient to cause us to forgo possible health benefits". On the contrary, the NBAC report spelt out many of the other ethical objections from both religious and secular sources, and stated its intention of setting the basis for a national discussion on such issues.

At last week's meeting, Kass tabled a simplistic tale of scientific hubris, *The Birthmark*, by the nineteenth-century writer Nathaniel Hawthorne (see www.vt.edu/vt98/academics/books/hawthorne/birthmark). That, and his support of human instinctive distaste as a fundamental moral measure of new developments, suggest a determination to confront the research agenda not only with ethical discussion but also with irrational fears and pessimistic foreboding. Who knows? If such a confrontation leads to a full and open debate, science may emerge more trusted, rather than less. But suggestions by the panel that there was little moral difference between reproductive and therapeutic cloning and that the word therapeutic might therefore not be used suggest that both rationality and language are threatened in this panel's deliberations.

Although perhaps tactically prudent at this early stage, it was worrying that no scientific organization took the opportunity to make a presentation at this open meeting. Scientists on the panel find themselves in a deck stacked with fundamental opposition. Those Americans, including scientists and ethicists, who hope to retain the freedom to pursue therapeutic human cloning, following a thorough assessment of the moral issues, should seek to ensure that any appeals to prejudice emanating from this panel are highlighted for what they are. ■

More speed, less speeding

An additional contents page in this week's issue reflects an extra service to which we are committed, but with provisos.

There have been many occasions in the past when this publication has released papers online ahead of their appearance in print. Usually these have been high-profile papers announced at a conference. In such cases, we felt it desirable or even essential to make the paper widely available as soon as possible.

In principle, the sooner a paper is published, the better for all concerned. What is more, researchers increasingly access papers central to their research via the web rather than print. So it is a natural and desirable step to make such advanced online publication a more routine process. This we have already initiated. The appearance this week of an additional printed contents page reflects our commitment to the service and an expectation that the number of papers published in this way will gradually increase. But although we recognize that competition needs to be taken into account, we will oppose the misuse of advanced online publication in that context, resisting

any tendency to cut corners from the process of publication.

The point at which publication occurs should be clearly and unambiguously defined: publication occurs at the moment the paper is first made available to *Nature's* readers, that is, when it is posted online, which could be on any day of the week. Moreover, the online version will be fully edited, and not subject to subsequent change.

Citations of print references are of little use to researchers accessing the literature online. Thus, we will encourage people to refer to a paper's Digital Object Identifier (DOI; see www.doi.org) before its appearance in print and, after printing, to retain a citation to its DOI alongside the traditional print reference.

Nature papers published ahead of print can be seen at www.nature.com/nature/aop, from which there is a link to a full explanation of the Nature Publishing Group's Advance Online Publication service (AOP). AOP is to be used by all of the *Nature* research journals. ■





Congo volcano
Early warning failed
to avert tragedy of
Goma eruption
p353



Lost property
Experts call for
reform of Japan's
patent system
p354



Burning issue
Expedition fired up
over hunt for
flammable ice
p355



Uncertain future
Cabinet reshuffle
leaves Canadian
science in limbo
p356

Panels' conflicting views cloud legal future of human cloning

Erika Check, Washington

With the US federal government deadlocked on legislation to ban or restrict the cloning of human embryos, several state governments are moving to pass their own laws on the issue.

If it continues, the flow of state legislation could result in a patchwork of rules across the United States, with 'therapeutic' cloning permitted in the more liberal coastal states, but banned in the conservative heartland.

The stalemate on the issue in Washington was neatly captured on 18 January, when two important protagonists in the debate presented strongly conflicting visions of appropriate rules for cloning.

On that day, a National Academy of Sciences panel, chaired by developmental biologist Irv Weissman of Stanford University, released the findings of an eight-month study. The panel recommended a legislative ban on reproductive cloning but urged the government to allow research on cloned human embryos.

For cloning research:
Irv Weissman
(centre) with
panel members
Maxine Singer
and Mark Siegler.



Meanwhile, a few blocks away in the Loews L'Enfant Plaza Hotel, the President's Council on Bioethics held its first meeting—and heard several of its members energeti-

cally denounce the use of cloning for research purposes. Indeed, the council's chair, Leon Kass, a bioethicist from the University of Chicago, declined to use the term 'therapeutic cloning', after council members objected to it.

"I do not believe there is a distinction between reproductive and therapeutic cloning. I think the distinction itself is morally problematic," said council member Robert George, director of the James Madison Program in American Ideals and Institutions at Princeton University.

The president's 18-member council consists of 15 academics—including 3 scientists—a journalist and 2 clinicians. The scientists urged the others to consider the potential therapeutic benefits of research cloning. "The question of how important this research may be to medicine is totally unresolved, which is a reason why we must be cautious about not preventing American scientists from pursuing these questions," said Janet Rowley, a biologist at the University of Chicago.

Kass himself noted that in the past he has



Leon Kass's council members object even to the term 'therapeutic cloning'

Geneticist takes top job at museum

Sally Goodman, Paris

Members of staff at the National Museum of Natural History in Paris are hoping that the appointment of a new president on 14 January will put an end to years of internal crisis there.

Bernard Chevassus-au-Louis, a specialist in the genetic improvement of farmed fish, comes from the French food safety agency, AFSSA, where he was president.

The museum has been managed by an interim administrator since September 1999, while a committee of external scientific advisers drew up plans to reform its research activities (see *Nature* 409, 273; 2001).

Chevassus-au-Louis is expected to reorganize the museum's 30 laboratories into a smaller number of departments around a theme of diversity in the living

world. "The museum needs a coherent global vision," he says.

The new president will be supported by a director responsible for the day-to-day running of the museum, and by a new scientific advisory committee. He will also appoint a director of collections to oversee a major planned renovation of specimen housing and cataloguing at the museum.

"The re-emphasis on the care of collections is good news for the world scientific community," says Peter Raven, director of the Missouri Botanical Garden in St Louis.

The government has promised 150 million euros (US\$133 million) over six years—less than the 400 million euros thought necessary for the renovation, but Chevassus-au-Louis believes the rest will follow.

DOUG MILLS/AP

M. SPENCER GREEN/AP

argued that the only way to enforce a ban on reproductive cloning would be to ban research cloning as well. He warned, however, that this did not necessarily mean that the panel would oppose the practice. "There is a division in this room about the feasibility and morality of doing research cloning, and where people will come out on that I don't know," he said.

Some observers emphasized the Bush council's strong conservative leanings. "This panel is heavily theological and religious and is also made up of many people whose first impulse about science and technology is sceptical," said Arthur Caplan, director of the Center for Bioethics at the University of Pennsylvania.

In any case, it remains unclear what role the panel will have in influencing federal cloning legislation. The House of Representatives has already supported a law that would ban cloning altogether, whereas the Senate is likely to favour a narrower ban that would allow therapeutic cloning.

With neither body inclined to budge, observers say that Bush may have little to gain from taking a strong position on the issue, whatever his bioethics council recommends. Bush has already expressed a desire to see all types of cloning outlawed, and his council may well adopt the same position. But this does not necessarily mean that he will take the politically risky path of trying to force the Senate's hand.

As the federal government considers its position on the issue, science marches on. In November, the Massachusetts biotechnology firm Advanced Cell Technology claimed that it had cloned human embryos, prompting many states to draft their own cloning laws.

Last week, for example, an advisory council weighed in on California's ban on reproductive cloning, which is set to expire at the end of this year. The advisory council, chaired by Hank Greely, a bioethicist at Stanford University, recommended that California renew the ban, but also said that the state should continue to allow therapeutic cloning. The state legislature will soon be considering whether to make all cloning illegal or simply to outlaw reproductive cloning and leave research cloning untouched.

Across the country in Wisconsin, the state in which James Thomson first isolated human embryonic stem cells, legislators are evaluating several proposals to restrict cloning. State houses in Massachusetts, Kentucky and Colorado are also debating their own laws.

But a Florida proposal is perhaps the most prophetic. The bill bans cloning for any purpose, and adds a uniquely American twist by allowing a cloned human to sue the scientists involved in his or her creation. ■

Craig Venter leaves top Celera post

Colin Macilwain, Washington

Craig Venter, founder of Celera, is stepping down as president of the Rockville, Maryland-based biotechnology company, which last year published a version of the human genome sequence (*Science* 291, 1304–1351; 2001).

Tony White, chairman and chief executive of Celera's parent corporation, Applera, will stand in as president for Celera while a new president is sought, according to a statement issued by Applera on 22 January. Venter, who was travelling and

unavailable for comment, will remain chair of Celera's scientific advisory board. But the statement said he will spend more time on his duties as chair of The Institute for Genomic Research, the Rockville company operated by his wife, Claire Fraser.

Although Venter was widely credited with Celera's success in genome sequencing, doubts have been expressed about the company's long-term business model, which has switched emphasis from users paying for data access to drug development.

Japanese labs could merge in drive for sharper focus

David Cyranoski, Tokyo

Moves are afoot to merge several of Japan's leading government laboratories into a single structure somewhat similar to Germany's Max Planck Society.

Advocates of the merger say that it would provide a focal point for top-quality research, and would aid collaboration between laboratories specializing in different disciplines. But critics are already preparing to resist the plan, which will surface later this month in a discussion paper to be published by the education ministry.

The plan would bring together some 14 government-run research institutes, including the High Energy Accelerator Research Organization (KEK) in Tsukuba, north of Tokyo, the National Institute of Genetics in Mishima, south of the capital, and the National Museum of Ethnology in Osaka.

Proponents of the reform in the education ministry are hoping to build a structure that could allow quick entry into new fields of research, close down outdated research facilities, rationalize administration and incorporate a more systematic approach to intellectual property (see page 354). They say that the new structure would give the laboratories a stronger footing to survive the Japanese government's continuing efforts to streamline all of its activities.

But the proposal is getting mixed reviews. Yoshiaki Hotta, director of the National Institute of Genetics, points out that the merged group would still not be strong enough to be a central core for Japanese scientific research unless other major research institutes, such as the Institute of Physical and Chemical Research (RIKEN), were included. "If it could give scientists more of a voice in the government, it would be good," says Hotta, "but it's not clear that that will happen."

Other institute heads say they fear that the merger would force them into negotiat-

ing their research programmes with other, unrelated institutes, instead of directly with the education ministry, as happens at present. "These institutes have such different purposes," observes KEK's director, Hiro-taka Sugawara. "How can they have a consistent organization that functions as one?"

But Hideo Mohri, president of Okazaki National Research Institutes (ONRI) in Aichi, another institute included in the plan, is enthusiastic. When the ONRI was formed from three institutes, they each retained autonomy, he says. "A merger is not a bad direction to go in," Mohri says.

Most agree that a merger could create opportunities to share resources. KEK's X-ray facility could prove useful, for example, to archaeologists at the National Museum of Japanese History who want to identify artefacts, says Sugawara.

"It could work, but many things still need to be ironed out," says Hotta. Details about how to split the financial resources between institutes with budgets of widely varying sizes are still under discussion.

If it is approved by the education ministry and the Council for Science and Technology Policy, the reform plan could be implemented by 2004, government officials say. ■



The High Energy Accelerator Research Organization (KEK) may merge with other labs.

Seismic rumbling foretold Congo eruption

Tom Clarke, London

Two seismological stations on Mount Nyiragongo in eastern Congo gave several days' advance warning of the volcano's possible eruption, scientists working in the area say.

But the lack of a functioning government in the war-torn region may have prevented the evacuation of the nearby city of Goma, where 45 were confirmed dead and an existing humanitarian crisis was worsened by the eruption.

For the past decade, a Japanese team has sought to maintain a seismic network at the volcano. In 1994, five monitoring units were donated by the US Geological Survey's Volcano Disaster Assistance Program (VDAP) to form the Goma Volcano Observatory.

But fighting in the area and looting of equipment by armed militia camped on the volcano itself regularly forced volcanologists to flee, according to Jacques Durieux, director of the Active Volcanoes Study Group in Lyon, France, who last visited Mount Nyiragongo with VDAP scientists in 1994.

According to the VDAP, only two monitoring stations were working properly before the eruption. VDAP scientists nevertheless received a warning on 12 January — five days before the first eruption — of the possibility of “abnormal seismic activity”.

Durieux, who spoke with Congolese volcanologists on the day of the eruption, says that efforts were made to raise the alarm,

but the lack of governance in Goma makes it unlikely that any plan to evacuate the city could have been implemented. Local volcanologists “have no staff and no pay”, says Hawaii-based volcano consultant Jack Lockwood, who has accompanied Durieux on previous trips to Mount Nyiragongo.

Durieux now wants to improve the observation of Mount Nyiragongo and its neighbour, Mount Nyamuragira, some 14 km away, which is also active. But he says that “we can't expect developing countries to pay”.

Mount Nyiragongo last erupted in 1977, killing around 500 people, and threatened to do so again in June 1994, when some 800,000 refugees from Rwanda's civil war were camped on its slopes. Last week's eruption

was the largest in the volcano's history, according to geological evidence.

Predicting future eruptions of an active volcano such as Nyiragongo will require a far more elaborate monitoring system than the two stations currently in place, say VDAP scientists. The VDAP and other agencies have provided more comprehensive monitoring of volcanoes elsewhere in the developing world, but such monitoring is only carried out at the request of the host government. With the province of North Kivu, where Goma is situated, controlled by Rwandan-backed rebels, such a request was not forthcoming from the government of the Democratic Republic of Congo in Kinshasa, say VDAP officials.



Mount Nyiragongo in eastern Congo killed dozens in its biggest-ever eruption.

JACKY NAEGLER/REUTERS

UN attempts to boost biosafety in developing world

David Adam, London

A scheme launched last week will help up to 100 of the world's poorest nations to obtain basic technical knowledge about genetically modified crops.

The US\$38-million scheme, unveiled in Nairobi on 16 January by the United Nations Environment Programme (UNEP), is designed to help the countries protect their interests more effectively in negotiations with agricultural suppliers, the food industry and drug companies.

Advocates of the scheme say it will help to conserve biodiversity and foster economic development. But critics argue that the money will not go very far and that the increased knowledge may confuse countries about the risks of transgenic technology.

The three-year programme will centre on “building capacity for assessing risks, establishing adequate information systems and developing expert human resources in the field of biosafety”, says Christopher Briggs, manager of the project.

But with funds of only around

US\$400,000 per country — provided by the Global Environment Facility, a joint venture of the United Nations and the World Bank — Briggs concedes that this is just a “first step”. Follow-up funds of up to \$1.4 million will be available to countries that create the necessary administration, he says.

The initiative follows a trial scheme with several developing countries. In one of them, Namibia, officials report that the money will eventually enable local scientists to test for transgenic crops independently.

But the project's small scale makes it “like offering swimming lessons to people in the Sahara”, says Calestous Juma, director of the Science, Technology and Innovation Program at Harvard University, which focuses on the role of research in developing countries. He fears that resources may be diverted “from areas where the risks are known to areas where they are still being debated”.

Participation in the scheme should allow countries to adhere to the Cartagena Protocol on Biosafety, an international agreement on approaches to agricultural



Growing knowledge: the scheme could help farmers learn how to handle transgenic crops.

biotechnology (see *Nature* 403, 473–474; 2000). Under the protocol, exporters must identify genetically modified crops so that importers can judge the risks associated with them. The protocol comes into force when 50 countries ratify it. Only 10 have ratified it so far, but the UNEP programme is expected to encourage others to do so.

♦ www.unep.ch/biosafety

Z. MATHEWS/TOPHAM PICTUREPOINT/UNEP

Japanese forum urges rethink over patents

David Cyranoski, Tokyo

A group of leading experts in patents and technology transfer have called for a complete overhaul of Japan's intellectual-property system.

The National Forum for Intellectual Property Strategy, whose 11 members include business leaders, lawyers and academics, want reforms that will allow individual researchers to retain the rights to their inventions. The forum also says that Japan needs a patent office with greater technical capacity.

Hisamitsu Arai, a former head of the patent office and chairman of the forum, presented the 60-page reform plan to the Council for Science and Technology Policy (CSTP), the country's highest scientific policy-making group, on 15 January.

The plan's 100 proposals address intellectual-property issues in universities, industry and government. The group claims that its proposed changes could make Japan "the



Light compensation: the blue LED's inventor Shuji Nakamura was less than satisfied with his reward.

world's number one intellectual-property country" by 2010.

Supporters of the reforms cite an incident last year, in which two Japanese biologists were charged with industrial espionage in the United States after removing materials belonging to a former employer there, as

indicative of the low awareness of intellectual-property rights issues in Japan. Lack of such knowledge led to the incident, they say.

Arai told the council that inventors, whether from a university or company, should be awarded patents as individuals, and that the accrual of patents should carry as much weight as the publication of research in the professional evaluation of researchers. The forum also calls for Japan to get tough on enforcing intellectual-property rights in other Asian countries, particularly China.

According to Arai, who is chairman of Nippon Export and Investment Insurance, patents in Japan "are just jewellery, a costly decoration. They're expensive but they earn little profit on average."

Corporations tend to give the same low level of reward for any patent taken out by researchers, says Arai. He notes the unfairness of this for big money-making products such as the blue-light-emitting diode discovered by Shuji Nakamura while at the Nichia Corporation. Nakamura, who advised the forum, cited low compensation as part of his reason for moving from the Japanese company to a university in the United States (see *Nature* 412, 844; 2001).

Lack of trained manpower also dogs the Japanese patent system, the forum says. "There are no PhD-holders in Japan's patent office, against some 500 in the US patent office," says Koichi Sumikura, an expert in intellectual-property rules in biology at the National Graduate Institute for Policy Studies in Tokyo.

"The CSTP agrees with most of these proposals," says Hiroo Imura, a member of the council. "But some, such as the idea that individuals rather than research organizations should own the intellectual property, remain contentious." And Arai accepts that some of the proposals will be unpopular with academic researchers.

Nonetheless, patent reform is on the cards in Japan before long. The CSTP meets again this week to discuss the best way forward. "Intellectual property is a big topic now," says Imura.

Satellite set to keep track of whales

David Cyranoski, Tokyo

A Japanese research team is putting the finishing touches to a satellite-based whale-tracking system that looks set to fuel the international debate on whale conservation.

The system — the brainchild of Tomonao Hayashi, an engineer at the Chiba Institute of Technology — will use a satellite to probe the whereabouts of hundreds of whales.

Hayashi hopes the project will help to resolve the ongoing debate about Japan's whaling programme. Japanese researchers, mainly at the Institute for Cetacean Research (ICR) in Tokyo, say that whale stocks are sufficient to warrant hunting; the International Whaling Commission disagrees. To prove its point, Japan has been carrying out controversial 'research whaling', killing about 400 whales per year. "Both sides know too little about the whale's ecology to reach a conclusion," says Hayashi.

The project will initially involve just one or two whales until it proves its merit internationally, Hayashi says. The education ministry has invested ¥300 million (US\$2.3 million) in the plan, and the satellite will be launched as part of Japan's Advanced Earth Observing Satellite II (ADEOS-II) project.

Researchers will tag the whales by using an airgun to shoot a titanium pin into the blubber, a procedure that Hayashi describes as "minimally invasive". A probe, attached to the pin by a two-metre rope, will send signals to the satellite.

Apart from allowing monitoring of migratory patterns and whale acoustics, the



probes will record pressure and temperature, and take geomagnetic readings during whale dives, which can last for 10–30 minutes. Data will be relayed to the satellite whenever the whale surfaces. "Sperm whales are thought to dive 2,000–3,000 metres during this time, but no one could tell for sure," says Hayashi.

Hayashi is developing a special generator to recharge the probe's batteries — a propeller-like machine that will derive energy from the whale's motion through the water. Although the whales will probably be stuck with the pin for life, Hayashi plans to attach the rope and probe with an iron fastener that will eventually rust away.

Seiji Ohsumi, director-general of the ICR, says the research will be "epoch-making", adding that the project will help the institute to work on rational stock management.

Argentine crisis raises questions over plans for Australian reactor

Peter Pockley, Sydney
& Virginia Gewin, Washington

Plans for Australia's largest-ever scientific facility — a research reactor to be built at Lucas Heights near Sydney — could be thrown into disarray because of Argentina's economic crisis, critics of the project claim.

Environmental groups in both Australia and Argentina say that INVAP, the publicly owned Argentine company contracted to build the facility, will not receive sufficient funds from its cash-strapped government to allow it to complete the five-year project.

But INVAP's chief executive, Héctor Otheguy, says that the company is financially sound and that this month's devaluation of the peso (see *Nature* 415, 104; 2002) leaves INVAP in a stronger position. "As a company based on exports, the devaluation will be a plus, not a minus," he says, adding that the project is running on schedule. Costing A\$290 million (US\$150 million), the research reactor is scheduled to open in 2005.

Brendan Nelson, Australia's new science minister, agrees. "INVAP's activities under the contract are fully funded by the contract payments and are therefore not dependent on cash flow from the Argentine government," he says.

But Raúl Montenegro, president of the Environment Defense Foundation, a pressure group based in Argentina, claims that INVAP relies on the Argentine government for cash flow, which is now drying up. "INVAP has no capacity for self-funding," he says.

A decision to license the reactor's construction is due next month after the project has been examined by the Australian Radiation Protection and Nuclear Safety Agency. One aspect the agency must consider is the viability of the plan to dispose of the reactor's spent fuel, which would be reprocessed by INVAP and stored at an unspecified site in South Australia. ■



Building up: Australia's planned research reactor is its largest-ever scientific facility.

Expedition trawls sea bed for energy-rich gas crystals



Hot property: the team aboard the *Meteor* hopes to capitalize on the huge energy potential of methane hydrate (inset).

Quirin Schiermeier, Munich

An international team is this month scouring the floor of the Black Sea in an effort to gather more information on energy-rich methane hydrate crystals. Resembling ordinary ice, these structures are stuffed with natural gas and have been cited as a potentially massive source of energy.

Methane hydrate can form only at reasonably high pressures and low temperatures, and so is found mainly in the Arctic permafrost and on the sea bed along the edge of continental shelves. The volume of gas it contains gives it great potential as a fossil fuel — the US Department of Energy, for example, estimates that successfully tapping just 1% of existing methane hydrate could yield more energy than the world's entire reserves of natural gas.

Margasch, as the Black Sea expedition is known, is studying the structure and architecture of methane hydrate in sediments just below the sea bed. It plans to map their distribution and hopes to estimate the total quantity of hydrate available. It is also using video-guided tools to retrieve samples of the crystals.

"Its near-surface deposits make the Black Sea a preferable destination for studying hydrates," says expedition leader Gerhard Bohrmann, a geologist at the GEOMAR research centre in Kiel.

The team, which includes geophysicists,

geochemists, biologists, oceanographers and meteorologists from Germany, France, Ukraine and Russia, left Istanbul on 2 January on the German research vessel *Meteor*.

The expedition has so far made good progress, Bohrmann reports. The scientists have already surveyed an active underwater mud volcano that releases methane, he says, and they hope that the data gathered will help them to test theories of how the hydrates form. They are also studying organisms that flourish at depths where the water contains no oxygen, surviving instead on methane.

Meteorologists, meanwhile, are interested in the influence that the hydrates might have on the composition of the atmosphere. Methane hydrate could be a significant source of atmospheric methane, an important greenhouse gas. Natural releases of the gas into the atmosphere could influence climate change, and German researchers aboard the ship are trying to establish how these releases occur.

The breakdown of the hydrates, triggered by changes in water pressure and temperature, is also thought to be responsible for seafloor landslides and large water waves (tsunamis). But most experts confess to being sceptical of the popular theory that they also sink ships in the foaming waters of the Bermuda Triangle. ■

www.gashydrate.de/projekte/omega/margasch

Last smallpox stocks granted stay of execution

Washington The planned destruction of the last official stocks of smallpox virus was postponed indefinitely when the executive board of the World Health Organization (WHO) met in Geneva on 17 January. The stocks, held at research institutes in the United States and Russia, had been scheduled for destruction at the end of this year. But the WHO smallpox advisory committee said last October that more time was needed to study vaccines and develop treatment programmes.

Several nations that had previously advocated the destruction of the virus, including Brazil, India and Japan, supported the call for further research. The board's decision is expected to be endorsed in May by the WHO's governing body, the World Health Assembly.

Unknown quantities take science helm in Canada

Montreal Canada's science policy is effectively on hold after a cabinet reshuffle left two key positions in the hands of politicians whose views on science are not known.

Industry minister Brian Tobin, who had been regarded as a future prime minister, unexpectedly resigned from politics on 14 January, saying he wanted to spend more time with his family. Allan Rock, previously the health minister, will now take the industry portfolio, which includes responsibility for most science-related issues. Maurizio Bevilacqua replaces Gilbert Normand in the junior post of secretary of state for science, research and development.

Both Rock and Bevilacqua are yet to make clear their positions on key science issues such as public Internet access and the long-awaited policy document on innovation.



New Canadian industry minister Allan Rock has so far given scant indication of his policy plans.

French to use spare embryos for study...

Paris The French parliament is set to allow research on stem cells from embryos left over after *in vitro* fertilization treatment.

France's 1994 bioethics law prohibits research on embryonic tissue, but states that this legislation should be reviewed within five years. The bill to revise the law allows research on stem cells from surplus embryos, but not the cloning of embryos for therapeutic or reproductive purposes. It also creates a new agency to authorize human-embryo research. The lower parliament was due to vote on the bill as *Nature* went to press, and was expected to pass it.

The law is unlikely to come into force until at least 2003. In the meantime, research minister Roger-Gérard Schwartzberg wants researchers to be allowed to import tissue from surplus embryos.

...as British law stretches to make human cloning illegal

London The British government has won its appeal against a court ruling that threatened to foil its attempts to regulate human cloning.

The ProLife Alliance, a group opposed to human-cloning research, challenged British legislation in the High Court last November, arguing that existing laws did not cover embryos created by nuclear transfer (see *Nature* 414, 381; 2001). The court agreed, effectively making it legal to create a cloned human baby. The Court of Appeal has now overturned this verdict, ruling on 18 January that nuclear-transfer embryos are covered.

The new judgement is "admirably sensible", says Robert May, president of the Royal Society, adding that the legislation permits legitimate uses of cloning technology but outlaws reproductive cloning. Those hoping to use the technology to investigate the therapeutic potential of embryonic stem cells must apply for a licence from the Human Fertilisation and Embryology Authority.

Japan's ocean-drilling boat tests the water

Tokyo The Japanese ocean-drilling ship *Chikyu* — meaning 'Earth' — made its maiden voyage last Friday.

The vessel is a key component of the Integrated Ocean Drilling Program, a US-Japanese collaboration that is due to start in October 2003. It will focus on seismology and the search for bacteria beneath the Earth's surface.

Chikyu, which can sleep 150 researchers and crew, will drill to 7 kilometres below the sea floor, three times the current record. The ship will spend a year in dock for more construction before being equipped with its drilling and research apparatus. Its



Deep digger: *Chikyu* should smash the ocean drilling record when it enters service in 2006.

contribution to the drilling project should begin in 2006.

♦ www.iodp.org

Study of cancer sufferers will evaluate cannabis pain relief

London British research into the clinical use of cannabis is to be extended to include studies of pain relief in cancer sufferers. More than 100 people with terminal cancer will take part in the trials, in which whole-plant extracts of cannabis will be given as a spray under the tongue. The company behind the study, Salisbury-based GW Pharmaceuticals, is already testing cannabis on patients with multiple sclerosis and spinal-cord injuries.

The attitude of British officials towards medical cannabis use has softened in recent months. Home Secretary David Blunkett said in October that the law could be changed to allow clinical use if trials were successful.

♦ www.gwpharm.com

Britain says goodbye to foot-and-mouth

London Britain has finally been declared free of foot-and-mouth disease, 11 months after the first outbreaks of the cattle ailment.

Four million animals have been culled to contain the disease, which is estimated to have cost the farming and tourism industries around £5 billion (US\$7.2 billion). The outbreak also forced Prime Minister Tony Blair to postpone last spring's general election for a few weeks, and brought about the demise of the Ministry of Agriculture, Fisheries and Food (see *Nature* 414, 839; 2001).

The announcement comes after three months in which no new cases were recorded in Northumbria, the last British county to be declared free of the disease. It will be several weeks before the remaining restrictions on animal movements are lifted.

The political and scientific fallout from the epidemic is far from over. Several reports, including one published this week by the National Farmers Union, criticize the government's handling of the outbreak. A Royal Society study of infectious livestock disease will report its findings in the summer.

♦ www.nfu.co.uk

India online



Home computers: a range of technological projects aims to bring the Internet revolution to India's poor rural communities.

A limited telephone network and low levels of literacy make it difficult to get information technology into India's rural areas. But as K. S. Jayaraman finds out, the country's engineers have developed some innovative solutions.

Bangalore may be the hub of a burgeoning information-technology industry, but the information super-highway peters out long before it reaches the villages of rural India. More than 80% of Indian villages have no telephone connection. Few villagers can afford computers where connections do exist, and low levels of literacy prevent many from using any machines they may have access to.

This is unfortunate, as villages have much to gain from the information revolution. Affordable Internet access could transform the way local farmers deal with wholesalers, or could supply health workers with reliable, up-to-date information. But because such ventures generate little or no profit, the big telecommunications operators and computer firms are unwilling to invest in them.

But India's problems are prompting Indian solutions. Engineers at one of the country's top research institutes are using radio links to get villages online. Others have developed an affordable computer that operates in three of India's most widely spoken languages. And both systems are designed to be attractive to local entrepreneurs and community organizations, thus bypassing the need for external investment.

"The big telecoms companies are ignoring the rural areas where 70% of Indians live," says P. G. Ponnappa, chief executive of telecoms company n-Logue, based in Chennai (formerly Madras). "Our aim is to bridge this digital divide."

It is easy to see why telecoms operators are reluctant to invest. New connections typically

cost around US\$800 to install, and companies need a certain monthly return to justify such an investment. "Telecom networks are designed for people who can afford a \$30–40 monthly bill. Only 1.6% of Indians can pay this much," says Ashok Jhunjhunwala, head of an electrical-engineering group at the Indian Institute of Technology (IIT) in Chennai.

The extra mile

For around a decade, Jhunjhunwala's group has been working on cheaper systems, focusing on alternatives for the 'last-mile' connections between local exchanges and individual homes. Conventional networks use copper wires, which are expensive to install and maintain, and account for around three-quarters of the cost of the entire network.

In conjunction with Analog Devices, a signal-processing company based in Norwood, Massachusetts, Jhunjhunwala's team has developed a wireless alternative, known as corDECT, which uses radio waves to link exchanges with homes. Existing telephone networks are used to carry signals over long distances. Once they reach the local exchange, however, the signals are digitized and beamed to homes and businesses, where a small, wall-mounted device is used to receive them. By splitting voice and Internet traffic at the local exchange, the system provides a separate phone line with each Internet connection.

The system can reach subscribers in a 10-kilometre radius of the transmitter, or up to 25 km with the help of a solar-powered relay station. IIT researchers are looking at taking the system into areas where there is no nearby

exchange by sending signals along the copper wires that run alongside the tracks of India's vast railway network. The system is normally used for voice communications, but includes spare wires that have not yet been used.

Several Indian companies, including n-Logue, a spin-off from the IIT, are already producing corDECT systems with connection speeds of 70 kilobits per second (kbps)—faster than most home connections in richer countries. Another manufacturer, Midas Communication Technologies of Chennai, says it will soon offer 380-kbps connections.

The technology is well-established, and is also being considered for use in Europe and the United States. But unlike in developed countries, its success will depend on small firms rather than big ones. Cable television, for example, has expanded rapidly thanks to local entrepreneurs. Broadcasters sell satel-



Ashok Jhunjhunwala has pioneered a cheap way to link homes to local telephone exchanges.

lite dishes to local businesses and leave the distribution — usually by cables slung between poles and trees — to local operators.

"We were inspired by the success of cable TV in India," says Ponnappa. "In 1992 there were no cable TV connections. Today 50 million households have cable TV because of local people's involvement." A similar system helped to seed the growth of private telephone kiosks, 650,000 of which have sprung up since 1987. In both cases, revenue is split between national companies and local operators.

A corDECT system typically consists of a central unit, based at the exchange, and around 20 transmitters. At a combined cost of around \$25,000, the system should be within the reach of local businesspeople. Ponnappa predicts that most operators — known as local service providers (LSPs) — will find 500–700 subscribers within a 25-km radius, making the investment worthwhile. So far, progress is good: n-Logue is currently signing up two to three new LSPs per month.

Besides this, n-Logue also works with local companies that are keen to provide Internet access to the surrounding population. In one project, a sugar factory in Nellikuppam in the southern state of Tamil Nadu has helped to fund 65 local connections. Sugar-cane farmers who supply the company can check their account details online, as well as accessing information on fertilizer and pesticide prices. What previously took the farmer a bus ride to the factory is now just a mouse click away.

corDECT is also being deployed by Media Lab Asia at the IIT campus, an Indian offshoot of the successful Media Lab at the Massachusetts Institute of Technology. The project will provide 1,000 connections and aims to show that villagers can make sustainable economic use of the technology. Ponnappa cites the example of a businessman in the state of Rajasthan who used a corDECT connection and a \$50 web camera to become the first professional photographer in his village.

In the large central-Indian state of Madhya Pradesh, local government is also using the technology. Five hundred villages already use government-funded connections to access land records or commodity prices, and to



Touching base: transmitters connect corDECT users up to 25 kilometres away to the central unit.

send complaints to regional officials.

Despite corDECT's potential, one major barrier — poor literacy — could still prevent millions of Indians from accessing the Internet. India has 18 officially recognized languages and 10 different scripts. As less than 0.5% of the rural population can read and write English, special software is needed to adapt keyboards to the different scripts. Even so, around 48% of India's population cannot read or write their native language.

Writing wrongs

A pocketbook-sized machine dubbed the 'Simputer' could be the answer. Developed at the Indian Institute of Science (IISc) in Bangalore, the Simputer can be used to play sound files, send and receive e-mails and browse the Internet. A stylus and touch-sensitive screen are used to input information and an Internet connection can be provided through a modem or a corDECT digital receiver — although the Simputer can only read web pages written specially for it.

Unlike rival machines, the Simputer can be used by those who cannot read or write. Speech-synthesis software developed at the IISc converts text in downloaded pages into spoken words in three Indian languages — Hindi, Kannada and Tamil. Commands can be inputted by tapping icons on the screen and listening to the Simputer.

The driving force behind the Simputer is a desire to get reliable and affordable information-technology hardware into rural areas. "We gave our time for free," says Swami Manohar, an IISc scientist who worked on the project. "The aim was to promote the Simputer not as an end-product but as an evolving platform for social change."

Costs have been kept down by using the 'open-source' Linux operating system and inexpensive memory units and processors. When mass-produced, the Simputer should sell for around \$200. This is still too much for



Hands on: getting to grips with the Simputer, which does not require users to read or write.

many rural people, but machines can be shared by several users, each of whom can use a smart card to save their personal settings.

The developers hope that rural communities will buy one or more Simputers and make them available to local people in schools or community centres. They envisage health workers using the devices to look up medical information, and local people using it to communicate by e-mail. Midas director Shirish Purohit suggests that Simputers could also be used in place of conventional computers in Internet kiosks.

PicoPeta Simputers, a Bangalore-based company formed to manufacture the devices, says it should have a marketable product within a few months. Before then, the Simputer will be tested by schoolchildren in the central-Indian state of Chhattisgarh, who will use educational software beamed directly from a satellite above India. If it is successful, the project could be repeated in other countries in the region, say the project's organizers, the South Asia Foundation, a Paris-based charity.

Meanwhile, corDECT is expanding beyond India's villages. Two national operators are deploying the technology in Indian cities. And late last year, the first of 1,000 subscribers in the Paragomina municipality in the Brazilian Amazon went online. Companies in countries from Fiji to Iran are also importing the technology. "A total of 100,000 corDECT lines are in operation," says Purohit. "Multinational companies will take note of us very soon." The technology could be used anywhere, but Ponnappa says that it offers the biggest advantages in countries with poorly developed telecoms infrastructure.

Whether or not corDECT gains international acceptance, India's rural population looks set to benefit from it. Factor in the Simputer, and the dream of giving rural people access to reliable, affordable information technology looks realistic. The engineers-turned-entrepreneurs who have pioneered the new technologies certainly think so. "The Internet revolution," says Jhunjhunwala, "will hopefully be one that rural India will not be condemned to stand by and watch."

K. S. Jayaraman writes for Nature from New Delhi.

► www.tenet.res.in/cordec/cordec.html

► www.simputer.org

► www.picopeta.com



In the state of Tamil Nadu, local businesses are helping to fund public Internet facilities.

Making waves

Pulses of energy called planetary waves traverse the globe, protecting the Arctic ozone layer and influencing weather and climate. Atmospheric scientists are now realizing just how important they are, says Larry O'Hanlon.

At their largest scale, they straddle the Earth. Without them, ozone depletion in the stratosphere would be worse and more frequent than it is now. They influence the weather that we experience, and are thought to be an important missing link in climate models. They are planetary waves: mammoth pulses of energy created when wind streams crash into mountains or collide with the steep temperature gradients in the air above coastlines.

These atmospheric Titans are certainly starting to make waves at scientific conferences. At meetings of the American Geophysical Union, for instance, the amount of time given over to planetary waves and related issues has tripled over the past three years. The subtleties of their behaviour remain mysterious, but if researchers can crack the waves' secrets, a better understanding of weather, climate and ozone depletion should follow. "There's no question that it's a hot topic," says



Hot link: Paul Newman has helped to confirm the effects of planetary waves on temperature.



Planetary waves help to dispel the polar stratospheric clouds that contribute to ozone depletion.

Alex Hall, who studies climate dynamics at the University of California, Los Angeles.

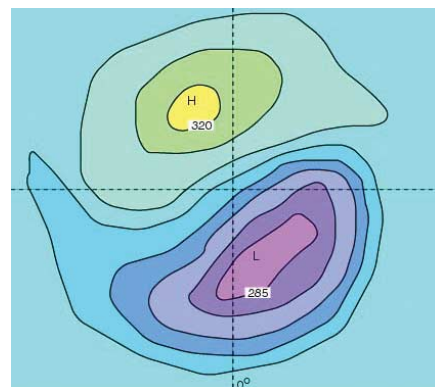
The Earth's atmosphere is full of waves, the smaller varieties of which provide our day-to-day weather. A ridge of high pressure that brings us sunshine for a day or two, for instance, is a wave — albeit a small one on the global scale.

Planetary waves reveal themselves over longer time and distance scales. For example, the high- and low-pressure fronts of the Northern Hemisphere are embedded in a giant wave. This has a low-pressure trough over central Asia and a high-pressure peak over North America — forming a huge 'yin-and-yang' pattern on plots of pressure centred on the North Pole (see right). Analyses of weather maps in the Northern and Southern Hemispheres reveal other such waves, some of which move across the Earth's surface.

Under pressure

Natural barriers to wind flow, such as the Rocky Mountains in North America, are one cause of such waves. Atmospheric pressure increases as the air piles up on the western side of the Rockies, and decreases as it expands after passing over the range. This creates a standing wave, fixed with respect to the Earth's surface. The wave is analogous to the patterns of water flow that form over rocks submerged in fast-moving streams, explains Paul Newman, an atmospheric physicist at NASA's Goddard Space Flight Center in Greenbelt, Maryland.

Some of the waves formed at mountain ranges move away from their sources. In the case of the Rocky Mountains, the high-pressure western side of the range emits waves of warm and less dense air, whereas cooler, denser waves leave the low-pressure eastern side. Some of the waves travel upwards, and the biggest — with wave-



Grand-scale wave: high pressure over North America and a corresponding low over Asia.

lengths of tens of thousands of kilometres — can pass up through the troposphere, the lower 10–30 kilometres of the Earth's atmosphere in which weather patterns move, and into the overlying stratosphere.

Even a perfectly smooth planet would create planetary waves, provided it had land and sea. Both air pressure and temperature change rapidly at the boundary between the two. Wind streams hitting this sharp gradient send reverberations of changing pressure, temperature and density outwards from the collision zone.

Planetary waves have attracted interest in recent years because of the way in which they influence the polar climate, a relationship that ultimately has a beneficial effect on polar ozone levels. Ozone in both the Arctic and Antarctic is destroyed because the low temperatures in the upper atmosphere allow the formation of clouds containing crystals of nitric acid, which absorb stable chlorine compounds throughout the winter. The spring sunshine and warmth releases these chlorine compounds, which are further broken down by sunlight into highly reactive

Understanding the effects of planetary waves could improve long-range weather forecasts.

chlorine molecules. It is these that go on to destroy ozone. In addition, ozone destruction is normally kept in check by nitrogen dioxide (NO_2), which mops up some of the chlorine molecules. But the nitric acid in polar clouds is formed from NO_2 , which effectively removes it from circulation.

Crystals of nitric acid only form if winter temperatures in the stratosphere fall below 1.78°C , and planetary waves limit this by raising the temperature. They do this in two ways, first by disrupting polar vortices — circular patterns of wind in the troposphere and stratosphere that isolate the poles from surrounding weather patterns, driving temperatures down. By disrupting the stratospheric vortex, incoming planetary waves allow warmer air from temperate latitudes to mix with and warm the polar stratosphere.

Upward-moving planetary waves also help to protect the polar ozone by 'breaking' when they enter the thinner stratosphere, much as water waves collapse and break when they hit shallower water. The breaking waves transfer their energy in the form of turbulence and heating to gases in the polar stratosphere.

Last September, Newman's team confirmed that the amount of energy deposited into the stratosphere by planetary waves in winter correlates with the subsequent temperature of the stratosphere — showing that



Acting as barriers to wind flow, the Rocky Mountains are the source of a major planetary wave.

the two are part of the same phenomenon¹. "For the first time, we showed the relationship," he says.

It has also long been suspected that upward-moving waves hitting the polar stratosphere drive the flow of ozone from the tropical stratosphere to the poles, which helps to top up polar levels².

But planetary waves benefit the two hemispheres to different extents. In the Northern Hemisphere, large land masses and high mountain ranges create high-amplitude waves. The south, with fewer mountains and smaller land masses, creates weaker waves and so gets less help. As a result, ozone over the Antarctic is largely wiped out for several months every spring, whereas only patchy holes form over the Arctic.

The Southern Hemisphere provides a cautionary tale of what could happen to the Arctic ozone layer if the influence of the Northern Hemisphere's planetary waves weakens — and the bad news is that the build-up of greenhouse gases in the atmosphere could have just that effect.

Greenhouse gases such as carbon dioxide and methane absorb various wavelengths of radiation and re-emit them as heat in all directions, ensuring that at least some of the heat is retained in the atmosphere rather than escaping into space. As heat radiated from the Earth's surface is trapped by the gases in the troposphere, less escapes into the overlying stratosphere. So, while the troposphere warms, the stratosphere cools³.

This stratospheric cooling could cripple the ability of planetary waves to mitigate ozone destruction. Currently, it does not take much energy from a planetary wave to raise the stratosphere's temperature above the threshold for the formation of nitric acid

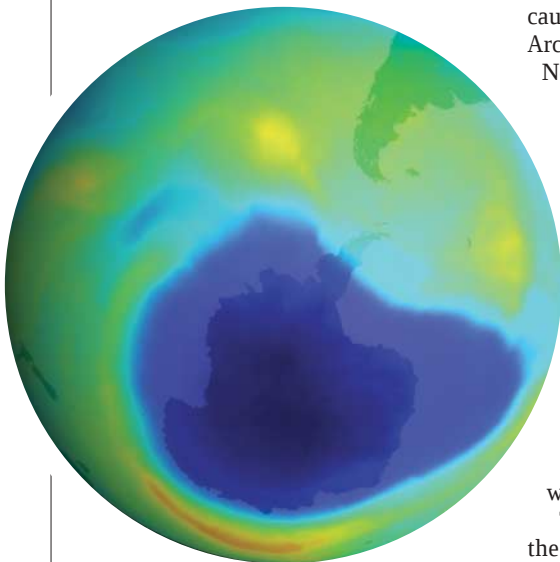
crystals. But in 1998, Drew Shindell and his fellow atmospheric modellers at NASA's Goddard Institute for Space Studies in New York showed that if the stratosphere gets colder, the planetary waves may not contain enough energy to raise the temperature sufficiently to shut the process down⁴.

A second blow comes from the effect of greenhouse warming on the westerly winds that stream across the Northern Hemisphere. These drive weather systems and feed into and reinforce the Arctic vortex. Together, the winds and the vortex form part of the Arctic Oscillation (AO). When the AO is in its positive phase, which can last for months or even years, barometric pressure drops over the polar regions and goes up at latitudes of around 55 degrees.

A tricky phase

Positive phases of the AO are more likely to occur during the winter, and the bigger the difference in atmospheric pressure between the pole and the lower latitudes, the stronger the westerly winds. The pressure differences have been getting bigger since the 1960s (ref. 5) and, although the reasons are unclear, greenhouse warming is a prime suspect. Whatever the cause, stronger positive phases of the AO could be a problem for the ozone layer, as the more powerful westerly winds that result would bolster the Arctic vortex and deflect planetary waves away from pole. If the pressure difference across the AO continues to increase over the years, the ability of planetary waves to rescue Arctic ozone will decline⁶.

Planetary waves may also be influencing the polar vortex and the AO in more subtle ways. Last October, Mark Baldwin and Timothy Dunkerton, meteorologists with the research firm Northwest Research Associates based in Bellevue, Washington state, showed that the stratosphere could influence the AO



Future shock: could the ozone hole above Antarctica (blue) be mirrored in the Arctic?



Twister: events in the stratosphere could provide clues to forthcoming freak weather conditions.

(ref. 7). They noticed that a weakening of the AO-strengthened vortex in the troposphere tends to be preceded, about two weeks earlier, by a weakening of the vortex above it in the stratosphere. The result was intriguing, as the low-pressure and virtually weather-free stratosphere was previously thought to have little impact on weather patterns below it.

Quite how this feat is accomplished is not known, but David Thompson, a meteorologist at Colorado State University in Fort Collins, suggests that planetary waves, one of the few messengers to cross between the two regions, could play a role. “No one understands this yet,” he says — but there is a lot of speculation.

Chain reaction

The chain of events may start when a planetary wave reaches the stable and non-turbulent airflow in the stratosphere. As the wave breaks, it transfers its energy to molecules in the stratosphere and deposits some tropospheric air. But on crossing into the stratosphere, some of the wave's energy is deflected in a way similar to how light is refracted when it passes from water to air, explains Thompson. The direction of the deflection depends on the stratosphere's temperature, and the direction and speed of its airflow.

Now comes the most speculative part, says Thompson: each wave that breaks within the stratosphere causes a disturbance and slows the stratospheric airflow. This, in turn, progressively alters the degree of deflection encountered by subsequent waves. That alone would seem to be a one-way, upwards effect. But the deflection also changes the airflow below the stratosphere in a process analogous to the way in which waters upstream of an obstruction in a stream can back up and slow

down, as if they are ‘feeling’ the change ahead.

Another possible way in which the lightweight stratosphere may influence the troposphere does not involve planetary waves, but simply the moving about of stratospheric air — subtly lightening the load above the troposphere in some areas and weighing down over others. It is not a lot of air or mass, but under some conditions it might be enough to tickle the troposphere into responding in kind and even switching between positive and negative AO modes, says Thompson.

A better understanding of the link could help weather forecasters to make improved long-range forecasts. Baldwin and Dunkerton's work suggests that changes to the weather in the troposphere could be foreseen weeks ahead of time by monitoring the behaviour of the stratosphere.

Baldwin has recently explored this connection with Wallace and Thompson⁸. “We wanted to see generally what was the scale of these effects on the hemisphere,” says Baldwin. The researchers also wanted to make clear the importance of the stratospheric connection with the weather, and to compare its effects to the temperature and pressure fluctuations of the El Niño/Southern Oscillation in the tropical Pacific Ocean.

They found that a pronounced weakening in winter of the Arctic's stratospheric polar vortex was often followed for about two months by an unusual increase in the number of intense winter storms and severe cold outbreaks in North America, northern Europe and eastern Asia. But a strengthened wintertime vortex brought unusually warm surface temperatures to the same, highly populated areas. The weather signal, they found, was just as loud as that caused by El

Niño, and so monitoring the stratospheric polar vortex could potentially be as useful as studying El Niño in predicting severe weather events, says Baldwin.

This top-down weather signal has also caught the attention of climate modellers, who until now have mostly ignored the stratosphere. “It's the opposite of everything people have been learning for a decade,” says Shindell. Modellers usually assume that the stratosphere follows the cues of its bulky big brother below. To save time and money, they leave it out of their equations.

Predictive power

But Shindell and his colleagues are among the minority who have kept the stratosphere in their models. As a result, they have accurately reproduced the rising winter temperatures in the Northern Hemisphere over the past three decades⁴, in addition to showing the stronger stratospheric polar vortices that set the stage for ozone destruction. Their model also recreates the way in which the stronger westerly winds deflect planetary waves.

On a shorter time scale, meteorologists are already used to keeping an eye on the joint influence of planetary waves and the tropospheric Arctic vortex. Planetary waves can disrupt the tropospheric vortex to an even greater extent than that in the stratosphere. Once weakened by the waves, the tropospheric curtain of winds that rings the pole and holds in the cold winter air can leak. Great masses of cold air are let loose, which can slide south along continental interiors. These leaks cause bizarre winter weather in temperate regions, such as freak snowstorms in Dallas, says Thompson. The extra-cold air masses can also collide with warm moist air, triggering the extreme storms seen in areas such as North America's Tornado Alley, a north-south corridor centred on the lowland areas of the Mississippi and Ohio river basins.

With planetary waves having a hand in such a wide range of atmospheric processes, and with the details of their behaviour still elusive, much work remains to be done. But thanks to the upsurge of interest sparked by the discovery of their influence on Arctic ozone, researchers from across the spectrum of atmospheric science are now taking notice. For those involved, it is an exciting time. “A bunch of fields are coming together,” says Thompson. “It's all one big jigsaw puzzle.” ■

Larry O'Hanlon is a freelance writer in Cool, California.

1. Newman, P. A., Nash, E. R. & Rosenfield, J. E. *J. Geophys. Res.* **106**, 19999–20010 (2001).
2. Holton, J. R. *et al. Rev. Geophys.* **33**, 403–439 (1995).
3. Ramaswamy, V. & Bowen, M. M. *J. Geophys. Res.* **99**, 18909–18921 (1994).
4. Shindell, D. T., Rind, D. & Loneragan, P. *Nature* **392**, 589–592 (1998).
5. Gillett, N. P., Hegerl, G. C., Allen, M. R. & Stott, P. A. *Geophys. Res. Lett.* **27**, 993–996 (2000).
6. Thompson, D. W. J. & Wallace, J. M. *J. Clim.* **13**, 1000–1016 (2000).
7. Baldwin, M. P. & Dunkerton, T. J. *Science* **294**, 581–584 (2001).
8. Thompson, D. W. J., Baldwin, M. P. & Wallace, J. M. *J. Clim.* (in the press).

Basic flaws in US human research protection must be addressed

Adding rules and paperwork misses the point: supervision must be comprehensive and based on ethical principles.

Sir—The National Bioethics Advisory Commission (NBAC), of which one of us (H.T.S.) was chair, issued its last report, *Ethical and Policy Issues in Research Involving Human Participants* (<http://bioethics.georgetown.edu/nbac>; M.S. was project director) in August 2001. We had been asked by the Office of Science and Technology Policy under the Clinton administration to examine the supervision system for protection of human participants in research, in the light of concern about possible fundamental flaws in the system—as discussed in your Opinion article “Time to cut regulations that protect only regulators” (*Nature* **414**, 379; 2001). Our report contains 30 recommendations for a wide range of essential changes at the national and local level.

Our main conclusion is that a single, independent office is needed to oversee a unified, comprehensive federal policy embodied in a single set of regulations and guidance that would apply to all types of research involving human participants, and including policy development and regulatory reform. This would improve protection, reduce unnecessary bureaucratic requirements, and enhance the supervision system's credibility by conveying a common set of ethical requirements that must be applied to all such research. Further, proper protection would be provided for children, prisoners and others in vulnerable circumstances. Merely adding more rules would not improve protection, and would lead to more paperwork.

None of the recent initiatives by the Department of Health and Human Services (DHHS), or its Office for Human Research Protections (OHRP), addresses the most fundamental flaws in the current system. For example, OHRP cannot force federal departments outside DHHS to use the new federal-wide assurance developed by OHRP because it does not have the authority. DHHS initiatives to improve reporting of adverse events cannot eliminate regulatory inconsistencies between DHHS and Food and Drug Administration (FDA) requirements, because—even though DHHS could require the FDA to revise its regulations—DHHS cannot unilaterally change that part of its regulations.

Most troubling is the ethically

indefensible situation that some human participants are protected by federal law and others are not. Only 17 federal departments or agencies (not including the FDA) have identical regulations, known as the common rule, requiring that the rights and welfare of research participants be protected. Some federal agencies sponsor human research but have not adopted the common rule, for example the Department of Labor, the Nuclear Regulatory Commission and the National Endowment for the Humanities. Although some institutions carrying out privately sponsored research do voluntarily protect research participants, there is an unknown amount of research in colleges and universities, fertility clinics, weight-loss clinics, work sites and businesses in which people are involved without their knowledge or informed consent, and without independent review.

Currently, no federal office has the authority to make government-wide rules to protect research participants, even for publicly sponsored research. Although the common rule was a reasonable way to create a uniform federal system of protection, it has not been modified since 1991, despite efforts. President Clinton, for example, directed federal agencies under the common rule to include protection when classified research is conducted. Despite three years of effort led by the Office for Protection from Research Risks (OPRR: the office that became OHRP), a presidential memorandum, and a challenge in the US District Court, the rule was not modified.

Some federal departments have modified their own set of common-rule

Costs increased for Mars Express, not just Beagle

Sir—Your News story “Europe's Mars mission to pay out for Beagle lander” (*Nature* **414**, 679; 2001) reported that the European Space Agency (ESA) “has been forced to allocate 36 million euros” (US\$32 million) to the UK-led Beagle 2 lander. This is misleading.

ESA's Science Programme Committee agreed on 4–5 December 2001 an increase (actually of 35.4 million euros) in the cost

regulations by issuing additional regulations or guidance. The Department of Veterans Affairs, for example, issued regulations for the treatment of research-related injuries in its medical centres. While laudatory to strengthen protection, this type of action can promote inconsistency and certainly undermines the unification the common rule was supposed to establish.

A supervision system can be successful only to the extent that those involved in human research recognize their ethical obligations to protect participants. Merely complying with regulations does not fulfil this duty. What is needed is a culture of concern and respect. Federal government and professional organizations must promote educational training in human research protection, certification for individuals and accreditation for institutions. NBAC recommends that everyone directly involved in human research know his or her ethical responsibilities and demonstrate his or her competence to conduct research ethically.

No supervision system can guarantee that research will be conducted without risks to participants. Nor can such a system prevent the recent tragedies at several well-known institutions. But a comprehensive, flexible system based on ethical principles and focused on ethically substantive requirements, rather than on obsessive documentation requirements, can ensure that tragedies are rare and the rights and welfare of participants are respected and protected. Until Congress and the executive branch of government acknowledge and address the three basic flaws of the system discussed here, the current supervision system will fail to adequately protect participants and will continue to frustrate institutions, investigators and Institutional Review Boards. Now is the time for change.

Harold T. Shapiro*, Marjorie A. Speers†

*Princeton University, 335 Wallace Hall, Princeton, New Jersey 08544, USA

†Association for the Accreditation of Human Research Protection Programs, Inc. (AAHRPP), 915 15th Street NW, Ninth Floor, Washington, DC 20005, USA

to completion of Mars Express, the spacecraft that will carry Beagle 2 to Mars but which will itself carry out a major survey of the planet. The only new resource directly attributable to Beagle 2 is the aseptic assembly facility, which is required to meet planetary-protection requirements and was approved at a cost of 3 million euros.

Of course, the overall technical development, integration and testing for Mars Express and Beagle 2 are closely linked; Beagle 2 is not simply added to the payload at the launch site. Accordingly,

ESA is monitoring carefully the development of Beagle 2 as its schedule needs to dovetail smoothly with Mars Express.

David Southwood

Director, Scientific Programme, ESA8, 10 Rue Mario-Nikis, F-75738 Paris cedex 15, France

Diversivory tactics in environmental debate

Sir — The irrelevant observations by Stuart Pimm and Jeff Harvey about Bjørn Lomborg's book *The Skeptical Environmentalist* (*Nature* **414**, 149–50, 2001) exemplify the unfortunate tendency of some environmental activists, when challenged with well-founded objections to the scientific validity of their alarmist claims about the state of the planet, to respond with such diversivory tactics as counting the number of footnotes cited by their critics, disparaging their critics' credentials and misrepresenting their views — everything, in short, but dealing honestly with the evidence.

In his book, for example, Lomborg presents a detailed analysis, based on widely accepted United Nations (UN) data, to show how the alarmists have been consistently wrong about the global population and food supply. Contrary to the predictions of impending catastrophe by Paul Ehrlich, Lester Brown and Pimm himself (who in 1998 told a meeting at the American Museum of Natural History that the world population might reach 40 billion by the end of the twentieth-first century), the data actually show that, over the past four decades, per capita food production has increased substantially, even in the developing world, while population growth has slowed so that the world population will level off at about 9 billion by mid-century.

Lomborg shows with devastating effect how the alarmists have been able to generate their forecasts only by extrapolating from very short-term or local negative trends, while disregarding the larger positive trends. Yet, instead of discussing this evidence, Pimm and Harvey simply attack Lomborg's credentials by attempting to associate him with the absurd view that an ever-growing population could be sustained for the next 7 billion years. Lomborg says no such thing, nor anything like it.

Similarly, Pimm and Harvey dismiss Lomborg's detailed, well-founded critique of exaggerated extinction-rate predictions with the ugly charge that this is morally equivalent to denying the Holocaust. Yet far from denying that loss of biodiversity and other environmental threats are

occurring, Lomborg plainly states that these are indeed serious problems. What he is criticizing is the habitual exaggeration and white lies that have become the common currency of environmentalist advocacy (see the Correspondence by A. Trewavas, *Nature* **414**, 581–582; 2001).

It is thus particularly ironic that Pimm and Harvey level the supercilious charge that Lomborg's book ("like a bad term paper", they say) cites secondary sources. Lomborg does indeed cite secondary sources — to provide examples of the exaggerations and distortions made by environmental activists. But to demonstrate the inaccuracy of these pronouncements, Lomborg cites primary sources such as UN data and articles in peer-reviewed scientific journals.

Lomborg's whole point is that the refusal of some environmental activists to deal honestly with the data harms the credibility of both environmental science and environmentalism. Pimm and Harvey in their review appear to have provided a further example in support of this thesis.

Stephen Budiansky

Black Sheep Farm, 14605 Chapel Lane, Leesburg, Virginia 20176, USA

Bioweapon agents: more access means more risk

Sir — In the News Review article "Bioweapons: Delivering death in the mail" (*Nature* **414**, 837–838; 2001), you quote Stanford biophysicist Steven Block as opposing restrictions on access to bioweapon agents, endorsing having "a lot of terrific [biomedical] scientists working on the problem" and forecasting "a 'molecular arms race' between bioweapons developers and biodefence specialists".

We believe that increasing the number of institutions and people with access to bioweapon agents will increase the likelihood of their release. We find the idea of a government-sponsored, large-scale multi-site "molecular arms race" frightening, and likely to lead to disaster. (See also Opinion and News, *Nature* **414**, 235 and 237–238; 2001.)

We recommend a moratorium on new permits for possession or transfer of bioweapon agents until the following measures are put in place: national or regional limits on numbers of laboratories permitted access to bioweapon agents; minimal containment requirements (BL-3); minimal security requirements (video surveillance, entry guards, dual-key locks, personnel screening; requirement that two people should be present during any work); inventory-reporting procedures;

and inspection and review procedures.

We believe that all laboratories without current permits, and/or not in compliance with the requirements above, should be required to transfer (under special permit) or destroy stored bioweapon agents.

This scheme would restrict access to bioweapon agents, but would not, in our view, unduly restrict biodefence-related research. Research in laboratories without access to bioweapon agents could be performed using non-pathogenic or moderately pathogenic species as simulants (for example using *Bacillus subtilis*, *B. thuringiensis* or *B. cereus* as simulants for *B. anthracis*), and/or through collaboration with laboratories permitted access.

This scheme is more restrictive than current official proposals. However, we believe that it represents the minimum required to provide a real increase in security. The National Institutes of Health has announced a major funding initiative for biodefence research. Without new restrictions, this initiative is likely to increase the number of institutions and people with access to bioweapon agents, and with training in production, handling and modification of these agents — and thus, perversely, to make us less secure.

Richard H. Ebricht*, Nancy D. Connell†

*Howard Hughes Medical Institute, Waksman Institute, Department of Chemistry, Rutgers University, Piscataway, New Jersey 08854, USA

†Center for Biodefence, Lourenco Center for the Study of Emerging and Re-emerging Pathogens, New Jersey Medical School, Department of Microbiology and Molecular Genetics, UMDNJ, Newark, New Jersey 07103, USA

How scientists can take the initiative in schools

Sir — As a schoolteacher, I am in my own way pursuing the agenda of your Opinion article "Educating future scientists" (*Nature* **414**, 1; 2001). I direct a programme, the science research initiative, in which science students aged 16 and above take part in experimental research, in collaboration with universities. They also write for and produce a national magazine, *N-lighten*, on contemporary science, and participate in a biennial conference to present their own discoveries and critically appraise those of others from schools, universities and industry.

In view of the declining numbers of graduates entering careers in science, I strongly endorse your view that the practitioners of science are ideally placed to enthuse students in schools with the excitement of modern discoveries and their implications.

Mo Afzal

The King's School, Canterbury, Kent CT1 2ES, UK

On the trail of two trials

Development of an AIDS vaccine will not be helped by duplicative trials.

John P. Moore

The need to develop an effective vaccine against HIV-1 is now probably the world's greatest single public-health problem. The global AIDS pandemic — 40 million people are estimated to be infected at present — will only ever be controlled by the use of a vaccine that significantly curtails transmission of the virus. Because of the vast cost involved and the organizational framework required, the scientific resources of the US government are the most critical component of international efforts to develop a vaccine. For the past 15 years or so, such efforts have been dominated by the US National Institutes of Health (NIH) and the US Department of Defense (DOD).

The US government's involvement in vaccine-trial programmes is also necessary to ensure that trials are always conducted safely and ethically. The need to apply high standards should be obvious, yet some US scientists are believed to be planning to use private or charitable funds to test controversial vaccines in developing countries, without proper safety testing in higher animals or prior evaluation in humans in the United States.

Whatever some scientists' desire to cut corners, evade the scrutiny of the US Food and Drug Administration and 'save time', the development of an HIV-1 vaccine is not a race for glory. When corners are cut, crashes can occur that force the safe as well as the cavalier drivers into the ditch. Ensuring federal government oversight means that the road to an effective vaccine, although long, will be safe for all — especially for those participating in vaccine trials.

Effective action?

But are the government's resources being applied in the most efficient and cost-effective manner? Unfortunately, the competitive mentality also applies to federal agencies — although without ever compromising ethical standards. It has long been known in the HIV-1 vaccine field that the NIH and the DOD regard each other as rivals, not collaborators. Both have conducted duplicate trials of many vaccine concepts: denatured gp160 virus proteins, monomeric gp120s, and live virus vectors have all been tested by both agencies, usually more-or-less simultaneously. It has seemed as if each agency has felt compelled to 'shadow' the other, to insure against the embarrassing outcome of a working vaccine candidate emerging that was sponsored by the other agency.

This situation has led to a massive waste



Competitive approach: the race to claim the prize of discovering a vaccine against AIDS stands a chance of disrupting, rather than helping, the overall research effort.

of money and other resources, and was made even worse by the advent of the International AIDS Vaccine Initiative (IAVI). Whatever is said in public, the rivalry between the NIH, the DOD and the IAVI is an open secret among scientists working on HIV-1 vaccine development — whether it is competition for clinical-trial sites or for funding of the more promising vaccine candidates or scientific programmes. Too often, rivalries have been fuelled by on- and off-the-record criticisms, usually unjustified, of the NIH in the media.

This state of affairs should not be allowed to continue. A coordinated effort could make so much more progress than one that is fractured by institutional rivalries. There will be more than enough credit to be shared out if an effective vaccine is eventually made.

Duplicated effort

The fractured nature of US vaccine programmes is fully revealed by the provisional plans of the NIH and the DOD to conduct duplicate phase III efficacy trials of a second-generation HIV-1 vaccine. Details were disclosed by Jon Cohen a few months ago (*Science* **293**, 1973; 2001) in an article that was, unfortunately, published on 14 September and consequently was overlooked given the horrific events of that week.

Both proposed trials will evaluate vaccines based on combining a recombinant canarypox/HIV-1 vaccine vector, made by

Aventis Pasteur, with a monomeric gp120 subunit protein, made by VaxGen, as a boosting antigen. There are some differences in the canarypox vectors involved, and in the particular gp120 protein to be used, but for all practical purposes, these trials are duplicative.

The NIH-sponsored trial will be conducted in the United States, the Caribbean and South America, involving 11,080 volunteers at a cost of between US\$60 million and \$80 million. Funding will come from the NIH's National Institute of Allergy and Infectious Diseases, through the Seattle-based HIV Vaccine Trials Network. The DOD's rival trial will be conducted in Thailand, using 15,800 volunteers at a cost of between US\$35 million and \$40 million, and will start this summer, at least six months before the NIH's trial can begin. A decision on whether to conduct the NIH trial will be made in the next few months, once the phase II trials have been fully evaluated.

Few independent scientists are optimistic that the canarypox/gp120 vaccine will succeed in preventing HIV-1 infection, or will have a sustained, beneficial impact on the course of disease in those individuals who become infected after vaccination. This vector only weakly induces cellular immunity — the current focus of the NIH-sponsored phase II trials is to find out if as few as 30% of the vaccine recipients can develop a measurable cellular immune

A. GRILLO/AP

response at any (not every) time point after vaccination. But even with the bar set so low for 'success' at the phase II level, the vaccine trials network may still not be able to find a way to lift the vaccine over it.

Moreover, the gp120 boosting-antigen does not induce a significant level of neutralizing antibodies against naturally circulating viruses. The limitations of gp120 proteins as immunogens have been well-known since 1993, yet they are still being used, solely because they are the only safety-tested HIV-1 proteins available in sufficient quantity. Monomeric gp120 proteins are now in phase III trials as solo antigens, both in North America and in Thailand, but a preliminary analysis last November revealed no evidence of efficacy. The probable failure of the gp120 concept cannot now be formally documented until November 2002, when the first of the two efficacy trials is unblinded. Unfortunately, this is several months after the decision about using these same gp120 proteins as boosting antigens in the canarypox vaccine trials.

It would surely be prudent to wait for the results of the gp120 efficacy trial before deciding whether to continue using these proteins in humans. Some argue that a boosting antigen is needed to enhance the effect of the canarypox vector. But, although there is some immunological basis for this view, why use a boosting antigen that is, for all practical purposes, inert? Why not use a Gag antigen that might more efficiently induce T-cell help and hence boost the immune system? Product availability must not dominate scientific rationale in decision-taking on this scale.

A difficult challenge

Recent studies in macaque monkeys with the simian immunodeficiency virus (SIV) analogue of the canarypox/gp120 vaccine show that gp120, either alone or as a boosting antigen, has no effect on the outcome on the viral challenge (R. Pal *et al.* *J. Virol.* **76**, 292–302; 2002). Indeed, the gp120 component was so ineffective that the gp120 recipients were pooled with the true control animals to increase the statistical power of subsequent comparative analyses. In the study, all recipients of recombinant canarypox became infected on challenge, but there were modest, short-term reductions in post-infection viral load and a slightly better preservation of CD4 T-cells in the vaccinees over a several-month period.

Given that the combination of canarypox vectors with gp120 is not impressive in human or macaque trials, why is there still such enthusiasm for the two efficacy trials? The main reasons seem to be that the vaccine trial networks regard it as a 'practice run' for when a better vaccine comes along, and that 'we might learn something'. Unfortunately, just what might be learned, and how, in this

If this turf war cannot be resolved by the agencies involved, maybe the White House should take action.

large and very expensive rehearsal has not been clearly defined. The NIH, in particular, is always under pressure, sometimes unfairly, from various interest groups about the scale of its investment in HIV-1 vaccines; it may in this case be responding to the need to be seen to be doing something on a large and noticeable scale. Yet the NIH is, and always has been, the engine-room of HIV vaccine development globally, even if it does not often receive the credit it deserves.

Once is enough

The animal studies mentioned above provide some support for further evaluation of the recombinant canarypox immunogen in humans, although not for the inclusion of the gp120-boosting component. It might be possible to learn whether a properly quantifiable cellular immune response could slow the rate of disease progression in infected vaccinees, at least in the short term. Any beneficial effect is likely to be marginal, however.

But why do the trial twice? Although duplication can mean rapid confirmation, surely there is a better way to make progress, given the resources required for two independent trials? The DOD's trial is likely to start earlier than that of the NIH, and it is the cheaper, so logically the NIH could support the DOD and abandon its own effort. Informal soundings suggest that DOD scientists would welcome the NIH's involvement. Some scientists in the NIH's trial network seem to believe that the DOD trial is under-funded, to the extent that the results will be under-analysed and a determination of 'correlates of protection' perhaps impossible. If so, all the more reason for the NIH network to provide resources to assist the DOD in the execution and analysis of a single, collaborative trial.

The recent announcement of a strategic link-up between the Merck HIV-1 vaccine-development programme and the NIH's vaccine trial network has major long-term significance. Many in the HIV-1 vaccine field believe that the Merck vaccine (DNA prime plus adenovirus vector boost) is an enormously more potent inducer of cellular immunity than the canarypox vector. Perhaps the NIH's vaccine trial network should now focus on this approach, leaving the DOD to study the earlier-generation canarypox vaccine, by analogy with the IAVT's extensive and wise investment in trials

of an MVA (modified vaccinia Ankara) vaccine, a more potent immunogen than canarypox, in the United Kingdom and Kenya. Alternatively, the DOD could leave HIV-1 vaccine development to the NIH and concentrate on bioterrorism defence instead.

The NIH's and DOD's leaders will no doubt be criticized whatever they do (see, for example, events recorded by Jon Cohen in his recent book *Shots in the Dark* — W. W. Norton, 2001). But they can now, if they choose, provide wise, statesmanlike guidance to the HIV-1 vaccine field if they place their programmes' narrow interests second to the need to save both volunteer cohorts and increasingly precious research dollars from being wasted in a turf war. To quote John Marburger, the White House's scientific adviser to the president (*Science* **294**, 1645; 2001): "If there are two agencies trying to do the same thing, I get them together and we work it out... There is nothing like OMB (the White House Office of Management and Budget) to help straighten out turf issues between agencies." If this particular turf war cannot be resolved by the agencies involved, perhaps the White House should take action. There are several possible options, but duplication is the least palatable.

Tragic price

Although there is increasing pressure on the AIDS research budget nowadays, even more important is the need to minimize the number of human volunteers used for testing vaccines that are highly unlikely to work in the traditional sense of preventing HIV-1 transmission. Two efficacy trials of the VaxGen gp120 product are now in progress, so conducting two more canarypox/gp120 trials would mean four consecutive, probably failing, efficacy trials, all conducted in the glare of extensive publicity.

The fear of failure should never dominate decision-taking, but when failure is probable, it is prudent to examine the likely consequences. The price of the failure of four trials in succession will be high — the erosion of public confidence in science's ability ever to deliver an effective HIV-1 vaccine. This is especially the case in developing countries, in which public-health officials have all too often been told that the present generation of vaccines will help their populations when, in all probability, they will not. If multiple vaccine-efficacy trials all fail, the consequent loss of confidence in the West's ability to stem the AIDS pandemic could have tragic consequences for the developing world. ■

John P. Moore is in the Department of Microbiology and Immunology, Weill Medical College of Cornell University, New York, New York 10021, USA.

Acknowledgements

I thank scientist and AIDS activist colleagues who contributed useful thoughts to this article.

Renouncing materialism

The fallout from 200 years of industrial creativity must be dispatched.

Materials Matter: Toward a Sustainable Materials Policy

by Kenneth Geiser

MIT Press: 2001. 317 pp. \$63, £49.50 (hbk), \$24.95, £20.50 (pbk)

Tim Jackson

It is a wet and windy afternoon in 1934, and a DuPont laboratory assistant is stirring a beaker containing a mixture of dibasic acids and diamines. As she absent-mindedly pulls the stirrer from the glass, the assistant notices that it trails a long, ductile fibre which soon hardens into a strong, silk-like thread. A few years later, the same assistant was probably clothed from head to toe in the stuff. DuPont called the fibre Nylon. It was just one of a myriad new industrial materials to emerge from a chemical industry rampant. From synthetic fibres to organic solvents, from polychlorinated biphenyls (PCBs) to chlorofluorocarbons (CFCs), from Teflon to Kevlar to a thin surface film called Tyvek, the twentieth century bore witness to an extraordinary proliferation of new materials, new uses for materials, new markets for materials — and new problems associated with them.

Kenneth Geiser documents this explosion of material ingenuity with grudging respect. But he makes it plain that his book is motivated more by concern than by professional admiration. Even as it was happening, there were those who bemoaned the exploding material lexicon of the new industrial age. 'Encase your legs in nylons,/Bestride your hills with pylons/O age without a soul!' wailed the British poet laureate John Betjeman in the postwar years.

By the end of the twentieth century it was abundantly clear that aesthetic losses were the least of our concerns. The fallout from 200 years of industrial creativity was a handful of toxic disasters, a host of minor calamities and a whole spectrum of environmental and health concerns that have, in retrospect, the aura of inevitability. We opened Pandora's box and paid little or no heed to the consequences. Short-sighted design criteria and endemic policy failure have bequeathed a legacy of toxicity and regret. Geiser is unequivocal in arguing that we must do better in the future.

This is not just hair-shirt environmentalism, however. True, the earlier part of the book is dedicated to the history of the material profligacy and policy failure that left us in this mess. But the second half is devoted to an exposition of the alternatives, an agenda for doing better, the guiding principles of a sustainable materials policy. Those familiar with the literature on resource productivity and



Expensive style: we are paying a high price for desirable modern accessories such as nylons.

reduction in the use of toxic chemicals will recognize at least some of those principles. Geiser highlights two main avenues for exploration: dematerialization and detoxification.

Dematerialization refers to a whole raft of measures — from improvements in the efficiency of procedures to materials recycling; from component reuse to the leasing of products; from new, radical ways of delivering services, to simply making do with less. One of the profound errors encoded in the institutional architecture of the modern economy is the linking of revenues (and hence profit) with the bulk throughput of material commodities. Dematerialization demands that we question that strategy. It requires that we restructure the economy in a variety of ways — some commonplace already, some radical in conception — that decouple well-being from material dependency. 'Living twice as well on half as much' is one of the epithets coined by the Factor Four movement — so called because it believes that a fourfold reduction in material intensity is not only possible but essential if we are to achieve sustainable development. Others have argued for factor 10 and even factor 100 reductions in material intensity. The important point is that this is a principle that should appeal even to narrow commercial interests: doing more with less — simple efficiency improvement — is, after all, one of the cornerstones of the market economy.

The fact that not everyone benefits commercially from this strategy — typically, materials consumers benefit and materials producers suffer — suggests that navigating



these efficiency improvements may not always be a painless process. One of my few reservations about this book is the limited attention paid to the institutional and social complexity of materials policies. Compensating for the social implications of massive industrial restructuring has never been a strong point in modern market economies. But if the scale of structural change envisaged by the dematerialization strategy is ever to be fully realized, social policy may turn out to be as important as industrial or economic policy in guiding the way.

The detoxification strategy — others have dubbed it the substitution principle — speaks (more or less) for itself. Since Rachel Carson's *Silent Spring* first alerted the world to the dangers of dissipating toxic chemicals into the environment, public and policy

DOUGLAS/ENG/AP

DUPONT

concern has escalated. PCBs, CFCs, dioxins, lead, mercury, cadmium: these materials and many others have already been the subject of regulatory scrutiny. Some have been successfully phased out. Geiser argues — and again he is not alone in this — that we should be attempting to substitute wholesale for certain kinds of toxic chemicals. The release of toxic compounds into the environment should occur only at rates that are comparable with natural biological processes. Persistent synthetic (non-naturally occurring) substances should not be released at all. We should engage in a continuing search for inherently safer substitutes.

Nowhere is it suggested that either of these strategies is straightforward to implement. On the contrary, the book recognizes that both present enormous challenges to modern industrial society. Geiser's point is that, until now, we have had no integrated policy framework even remotely capable of pushing us in the right direction. The 'ecological' economist Nicholas Georgescu-Roegen once remarked that the fourth law of thermodynamics ought to be that "matter matters too". Geiser reiterates this theme, but he does more than this. He shows why it matters, where it matters and to whom it matters; and he makes a pretty good stab at figuring out what to do about it. This book should be required reading for industrial designers, materials scientists, chemical engineers and environmental policy-makers everywhere. ■

Tim Jackson is at the Centre for Environmental Strategy, University of Surrey, Guildford GU2 7XH, UK.

Perchance to dream

The Dream Drugstore: Chemically Altered States of Consciousness

by J. Allan Hobson

MIT Press: 2001. \$27.95, £19.50

Brian Johnson

In their 'activation-synthesis' model of 1977, J. Allan Hobson and Robert McCarley described dreaming as a response by the brain's cortex to random firing by the pons, a structure in the brainstem that is responsible for REM (rapid eye movement) sleep. Hobson and McCarley proposed that these meaningless stimuli become organized in the same way that someone might perceive a coherent picture in an ink-blot psychological test. With this new understanding, they concluded that the attention lavished on dreams by psychoanalysts was an expensive waste of time.

Hobson has now updated and expanded his thinking, and has come up with the AIM (activation, input, modulation) model,



Nightmare neurons: psychedelic brain painting by John D. Woolsey, taken from the cover of the book.

which describes how brain activity changes during waking, non-REM and REM sleep. The 'drugstore' aspect of the book compares alterations in the balance of neurotransmitters released in the brainstem, which are responsible for the shift in consciousness from waking to dreaming, with changes in the brainstem that account for the cognitive and behavioural characteristics of mental illness and drug-induced states.

Two important current research issues challenge the AIM/activation-synthesis model. They stem from the work of two neurologically oriented psychoanalysts, Mark Solms and Sigmund Freud. Hobson addresses the two issues in his book. The first concerns Hobson's assertion that the pathways involving the release of the neurotransmitter acetylcholine, which originate in the pons and energize REM sleep, are actually responsible for dreaming. Solms has found extensive evidence in patients with brain lesions that it is the dopamine-releasing pathways originating in the midbrain and running through the forebrain to the cortex that provoke dreaming, whether in REM or non-REM sleep. In other words, Solms asserts that it is the midbrain, not the pons, that is the driver of dreams. This is a key issue, because if Solms is right, the same dopaminergic system that energizes our drives for food, water and sex also stimulates dreaming. This would go a long way towards explaining Freud's theory that the origin of dreams is an often-disguised wish.

Another controversy regards the meaning of dreams. Hobson now sees dreams as emotionally salient because he acknowledges the existence of inputs to the pons, via descending fibres, from the mood and memory areas. Therefore, he is halfway back to

agreeing with Freud. However, he explicitly denies that dreams have any latent or unconscious content: "It sure looks like we aren't supposed to remember our dreams. That can't be because our consciousness would be damaged by them, as Freud assumed, but it could be because they contain useless or even misleading information that is better ignored than taken too seriously."

Can he afford to claim that certain aspects of a dream convey meaning whereas others must be ignored? The book contains examples of Hobson's dreams and his interpretations. For example, he builds one of his chapters around a dream he had during the time he was writing about the AIM model. In the dream, Hobson's vacation retreat is being altered without his participation or control. It is becoming an architectural chimera — which, he assures us, he is *not* anxious about. He is unhappy about the fact that leaks are flooding his farmhouse, even though the pipes have been cut off *and capped* (his italics), and that his hospital's unique 1912 period furniture is being pilfered by his colleagues. As Hobson made associations between events at his farmhouse and in his lab, I found myself eagerly anticipating his interpretation of the associations to "pilfering". But at the end of the chapter he issues a denial: "I might be a scavenger, but I am NOT a thief or a looter, even in my dreams." This seems disingenuous. He describes many thoughts about every other part of the dream. The denial of unconscious or latent content regarding worries that the AIM model is springing leaks, or that something dishonest is going on, doesn't ring true.

Unfortunately, the uninformed reader will not realize how close Hobson has come to returning to a psychoanalytical model,

now bolstered by burgeoning neuroscience. Hobson never acknowledges his use of Freud's insights. On the contrary, he asserts: "By equating what he took to be the psychodynamics of dream formation with all neurotic and psychotic symptom formation, Freud built a house of cards that is now tumbling down." A few pages later, he claims as his own what many regard as Freud's insight: "This imaginative, autocreative aspect of dream consciousness may be worth studying ... as a way of understanding psychotic states, as I have often suggested." Hobson is full of dismissive comments about psychoanalysis — which is paradoxical, as he seems to be using analytical ideas. His combative stance flaws an otherwise interesting book. ■

Brian Johnson is at 5 Park Place, Newton, Massachusetts 02460, USA.

Making waves on rocky ground

Tsunami: The Underrated Hazard
by Edward Bryant
Cambridge University Press: 2001. 320 pp.
£19.95, \$27.95 (pbk)

Kenji Satake

A tsunami, one of the few internationally understood Japanese terms, is a water wave in an ocean or lake. It can be generated by various submarine geological processes, such as earthquakes, volcanic eruptions, landslides and even meteorite impacts. In the deep ocean, a tsunami travels at jet-liner speed without raising or lowering the sea surface by more than a few metres. In shallow water, however, the wave becomes slower and bigger, sometimes rising to more than 10 metres and causing extensive damage to coastlines.

In 1960, a great earthquake off the Chilean coast sent a tsunami to Hawaii and Japan that resulted in many fatalities. Such waves threaten not only the Pacific Rim but also the shores of the Atlantic Ocean and the Mediterranean and Caribbean seas. In the past decade alone, some 5,000 deaths have been caused by 10 tsunami events.

Edward Bryant provides a wide-ranging review of tsunamis. His book starts with five stories of tsunamis, pieced together from legends, historical documents and eyewitness accounts. The book ends with another five scenarios of future potential tsunami damage around the globe. Although the latter collection verges on scientific fiction, it carries the author's message that tsunami hazards are underrated. The intervening chapters explain the nature and context of the tsunamis described in the stories. The book contains vivid descriptions of past tsunamis, and makes effective use of graphics, photographs and imaginative illustrations.



Waves of destruction: tsunami events have caused 5,000 deaths in the past decade.

Recent advances in tsunami science have come from many fields. Broad-band seismic records help seismologists to estimate the space-time distribution of tsunami-generating seafloor displacements. Coastal and ocean-bottom gauges now monitor the progress of tsunamis and improve the accuracy of tsunami warnings. Physical and numerical models simulate the generation of tsunamis and their interactions with coastal structures such as breakwaters. Fast computers help to identify areas that are at greatest risk from tsunamis, as does the identification of onshore deposits of past tsunamis. The Internet has increased the rate at which information on tsunami events and research can be exchanged. Large tsunamis attract international, multidisciplinary survey teams.

Bryant became interested in tsunamis when he was investigating the origin of boulders in Australian coastal rock platforms. The book, in both its scope and its wealth of citations, demonstrates that he has read widely. And the case studies and research results he cites are up to date. That alone makes the book useful as a reference work. However, in quality and depth, the book varies greatly from chapter to chapter.

The section headed "Tsunami-formed landscapes" will interest scientists studying prehistoric tsunamis or coastal geomorphology. These chapters deal with deposition and erosion resulting from tsunamis — a topic that has received attention only in the past decade or so. Bryant describes tsunami deposits worldwide, drawing particular attention to boulders he found in southeast Australia, atop cliffs and dunes tens of metres above the current sea level. He repeatedly distinguishes between tsunami deposits and storm deposits, although the criteria used for this are not always obvious. In fact, such criteria have been a hot research topic. A summary of storm deposits would have been a useful addition. Bryant reviews reports of tsunami deposits in Hawaii, although without due attention to attendant controversies on their origin.

Bryant also explores ways in which

tsunamis can shape coastal landscapes, particularly by 'scouring' bedrock. He thereby revives Sir James Hall's postulation of 1812 that Edinburgh's landscape was shaped by a tsunami, and draws an analogy to scouring by subglacial streams. This unique idea, so far untested, states that high-speed rotation of water creates a whirlpool which then scours rocky coastlines. But it requires a quantitative explanation involving factors such as angular velocity and energy source, whereas Bryant simply invokes cavitation. A modern example of bedrock scouring would also have made Bryant's arguments more convincing.

Other parts of the book are written for the general reader. Some chapters barely go beyond student-notebook level and often mix established theory with speculation. Although any separation of the two would be subjective, these sections need to be read with a critical eye. When discussing tsunami dynamics, Bryant presents equations without providing sufficient background on concepts such as the intrinsic difference between deep- and shallow-water waves or finite-difference methods. Also weak are the sections on seismic gaps and earthquake (and tsunami) prediction. And it would have been helpful to have described the conflicting arguments on earthquake predictability.

Bryant's book has lofty goals — summarizing a diverse and rapidly emerging field is not easy. The book succeeds in organizing and providing examples of many facets of tsunami studies, but falls short of portraying these facets accurately. It will therefore be most useful to tsunami researchers who can identify its technical and factual errors. Parts of the book lack rigour and consistency; especially needed are a more comprehensive and better-organized index and a glossary of tsunami jargon from the many fields involved. With these revisions and peer review, a second edition of this book could be an authoritative, multipurpose textbook for newcomers to tsunami research. ■

Kenji Satake is at the Active Fault Research Center, National Institute of Advanced Industrial Science and Technology, Tsukuba 305-8567, Japan.

When order comes naturally

Self-Organization in Biological Systems

by Scott Camazine, Jean-Louis Deneubourg, Nigel R. Franks, James Sneyd, Guy Theraulaz & Eric Bonabeau
Princeton University Press: 2001. 535 pp.
\$65, £43

Eshel Ben-Jacob

An old man is sitting by a pile of rocks, carving one into a block. A passer-by asks him what he is doing. "Can't you see?" says the old man. "I'm building a cathedral." What would a mound-building termite, carrying a ball of mud, reply? Does it know that it is taking part in the construction of one of nature's most astounding structures? The authors of *Self-Organization in Biological Systems* hold the belief that the termites in the colony don't have a notion of participating in a group task. The mound, they say, emerges through a process of self-organization, defined as "a property of certain dynamical mechanisms whereby structures, patterns and decisions appear at the global (colony) level of a system, based on interactions among its lower-level components (agents, individuals)". They favour the view that the process is executed on the basis of purely local information, without reference to the global level.

In the first part of the book, the authors describe self-organization and define relevant concepts, such as pattern, complexity and emergence. They explain how self-organization works, with emphasis on the role of positive and negative feedback, amplification of fluctuations, local information and information transfer between individuals. Self-organizing systems are characterized, modelling approaches discussed and alternative mechanisms presented.

Next, the authors address their primary goal: "to demonstrate, for a wide range of examples, the link between the rather simple behavioural programs of the individuals in a group and the sophisticated structures and patterns that emerge from their collective activity." With present computational power, it is only natural to use computer models to study self-organization. In physics and chemistry, models have long served as an indispensable research tool. But in the life sciences they have a different role, and many biologists still question the value and predictive power of models. This book should contribute to a recognition of the usefulness of modelling in the study of biocomplexity.

But simulations alone, without analysis, are insufficient. One can easily fall into the 'reminiscence trap', described by J. D. Cowan as the tendency to devise a set of rules that will



Group participation: worker ants joining leaves (top), and the spiral organization of a daisy head.



just mimic aspects of observed phenomena. But this can be prevented if one understands the fundamental mechanisms at their heart. Then, a model can serve to make reliable predictions, and not just reproduce existing behaviour. The desire to achieve this naturally pushes one in the direction of including more biological details. Nevertheless, the other extreme must also be avoided — the 'realistic trap' — where the model becomes swamped with too many details (and, usually, unknown parameters) and hence loses all predictive power. Model building, indeed, is "an art in its own right" (the skills can only be acquired by practice), the challenge being how to elicit the generic features and basic principles needed to explain behaviour from experimental observations and biological knowledge.

Many interesting examples of biological self-organization are presented — from complex patterning of bacterial colonies through mound-building termites and other social insects to schools of fish. In each of the case studies, the transition from observation to modelling is demonstrated, along with

model analysis, simulations and comparison with the observed phenomena, in which the authors take the reader, step-by-step, through the construction of the model.

Although local interactions are undoubtedly important, I don't agree with the authors that they are sufficient to explain bio-self-organization. For example, even simple microorganisms can gather information about the global state of their colony. It is known that bacteria chatter via chemical signals that propagate throughout the colony, thus providing each bacterium with global-level information. For example, some bacterial strains sporulate collectively when starved, but each bacterium does so only after broadcasting its starvation and receiving an echo — in the form of a chemical message — that the entire colony is ready to sporulate.

When studying self-organization and pattern formation in non-living systems, the questions asked are largely of a mechanistic nature. But living systems pose another question that is far more challenging — that of origin. The authors believe the evidence they have amassed in this book corroborates what has long been an automatic answer to this question — that "much of the complexity of self-organised structures seen in biology arises because the rules governing the interactions among the components of biological systems have evolved through natural selection". They raise this issue throughout the book, but I was not convinced, as on the basis of the very same information one may just as well arrive at the opposite conclusion — that self-organization and natural selection are two alternative evolutionary mechanisms. The authors do mention that some biologists hold the belief that the first is an alternative to the second, which they glibly dub a "misconception". But when the same set of facts can lead to two opposite conclusions, the reason might be that neither is really a conclusion. This reservation aside, the book unfolds, in a very interesting and useful manner, the body of work relating to the mechanisms involved in biological self-organization. ■

Eshel Ben-Jacob is in the School of Physics, Tel Aviv University, 69978 Tel Aviv, Israel.

More on self-organization

Emergence: The Connected Lives of Ants, Brains, Cities and Software

by Steven Johnson
Allen Lane/Scribner, £14.99/\$25

correction In the review by Howard P. Segal of *IBM and the Holocaust: The Strategic Alliance Between Nazi Germany and America's Most Powerful Corporation* by Edwin Black (*Nature* 411, 993–994; 2001), one paragraph stated: "Polish ghetto exterminators ... never had punch-card machines, and the SS Race and Settlement Office obtained them only in 1943." In fact, the SS Race and Settlement Office was an SS officers' marriage and adoption-screening bureau, not a prime mover in the extermination process.

The review also stated: "Black minimizes the inefficiency of many Nazi procedures, which, for example, delayed many trainloads of Jews en route to concentration camps." In fact, custom-designed IBM programs helped the Nazis to locate and deploy the locomotives and boxcars. Virtually all European trains were at that time tracked by Hollerith cards.

These small errors do not affect the rest of the review, or its general conclusions.

The physics of society

Philip Ball

The introduction of probability into the fundamental nature of the quantum world by Bohr, Born and Schrödinger in the 1920s famously perplexed some scholars of science's philosophical foundations. But arguments about chance, probability and determinism were no less heated in the mid-nineteenth century, when statistical ideas entered classical physics.

James Clerk Maxwell allowed probabilistic physics to bring him to the verge of mysticism: "It is the peculiar function of physical science to lead us to the confines of the incomprehensible, and to bid us behold and receive it in faith, till such time as the mystery shall open." The use of statistics as a mathematical tool of all the sciences provoked passionate responses from philosophers, novelists and social commentators. Few scientific issues besides darwinism (itself given a statistical treatment by Darwin's cousin Francis Galton) attracted such debate in parlours and periodicals.

It was not physicists who started this fuss but sociologists, who found that chance and randomness in the world of people and politics, far from banishing predictability and making social science oxymoronic, seemed to have laws of their own. This, to widespread dismay, seemed to challenge the idea of free will.

In the seventeenth century, Sir William Petty recognized that the study of society could only hope to emulate the precision of science if it became quantitative. Petty's call for a 'political arithmetic' induced his friend John Graunt in the 1660s to advocate 'social numbers' as a way to guide political policy. Graunt compiled mortality tables, reasoning that good legislation and government are impossible without such demographic data.

Births and deaths were a major preoccupation of early social statisticians, including the astronomer Edmund Halley. In 1781, Laplace tallied male and female births in Paris, explaining their near-equality as merely the result of a random process, rather than, as was previously thought, a sign of divine wisdom.

Laplace showed that variations in such social statistics could be described by a universal 'error curve', which was introduced in 1733 by the mathematician Abraham de Moivre to describe the results of coin tossing. The ubiquity of this curve, now familiar as the gaussian, was then seen as miraculous: a natural law that applies as much to human affairs as to errors in measuring planetary motion.

The idea that society is governed by laws as precise as those of physics was a product of the Enlightenment, and was espoused by Immanuel Kant and the political philosopher Auguste Comte. When the Belgian astronomer Adolphe Quetelet came to the French Royal Observatory in 1823, he was captivated by Laplace's statistical regularities and began to argue in favour of what Comte later called "social physics". Quetelet's popularization of Laplace's data impressed the likes of John Herschel and John Stuart Mill.

But the most visible exposition of these laws was given in the epic (and misnamed) *History of Civilization in England* (1857–61) by Henry Thomas Buckle, who believed that historical events occurred with a law-like inevitability. "The great truth," he said, "is that the actions of men ... are in reality never inconsistent, but however capricious they may appear only form part of one vast system of universal order." One of the book's earliest readers was Maxwell, who found it "bump-tious" but remarked that it contains "a great deal of actually original matter, the result of

Statistics

By seeking to uncover the rules of collective human activities, today's statistical physicists are cautiously aiming to return to their roots.

fertile study, and not mere brainspinning".

Others dismissed the idea that people's actions are governed by laws. At times, Buckle seemed to imply a compulsion whereby individuals would find themselves acting in a certain way to fulfil 'quotas'. Ralph Waldo Emerson mocked what he saw as the absurd rigidity of the idea: "Punch makes exactly one capital joke per week; and the journals contrive to furnish one good piece of news every day." In *Notes From the Underground*, Dostoevsky had Buckle in mind when his narrator raves that man would rather make himself mad than be constrained by law-like reason.

Maxwell concluded that the error curve is the signature of all random processes, and of processes that, although deterministic, are too complex to be reduced to newtonian terms. So the gaussian curve was the natural choice for the velocity distribution in his kinetic theory, and the statistics of social physics thus helped to launch statistical mechanics. Ludwig Boltzmann also drew on the analogy with the statistical regularities of Quetelet and Buckle when he extended Maxwell's work on molecular probability distributions in 1872.

There is a pleasing symmetry to the way in which today's statistical physicists cautiously seek to extend their models and concepts into social science, such as pedestrian and traffic movement and descriptions of the economy. Statistical methods are hereby returning, much refined, from whence they came.

The statistical nature of quantum mechanics is different from that of classical physics, as it invokes variables with values that are not merely unknown but unknowable. Nonetheless, quantum probability would have had a rockier path if physicists had not been prepared by the knowledge that a statistical approach does not preclude the existence of precise laws. As early as 1918, the physicist Marian Smoluchowski considered probability to be central to modern physics: "Only Lorentz's equations, electron theory, the energy law, and the principle of relativity have remained unaffected, but it is quite possible that in the course of time exact laws may even here be replaced by statistical regularities." ■

Philip Ball is a consultant editor of *Nature*.

FURTHER READING

Porter, T. M. *The Rise of Statistical Thinking 1820–1900* (Princeton Univ. Press, Princeton, 1986).
Olson, R. G. *Science Deified and Science Defied* Vol. 2 (Univ. California Press, Berkeley, 1995).

VICKY ASKEW



A molecular full nelson

Robert C. Liddington

The crystal structure of the final component of the three-part anthrax toxin shows how it disables a host signalling molecule, while simultaneously using that molecule to stimulate its own activity.

The infamous killer *Bacillus anthracis* would be a harmless, soil-dwelling bacterium were it not for two extra DNA molecules, set apart from its main chromosome, known as pXO1 and pXO2. pXO2 encodes the proteins that create a protective capsule for *B. anthracis*. pXO1 encodes, among other proteins, the three that are collectively known as 'anthrax toxin': protective antigen (PA), lethal factor (LF) and oedema factor (EF). On page 396 of this issue, Drum and colleagues¹ describe the crystal structure of EF in complex with its host-cell activator, calmodulin, and in so doing complete our atomic-resolution view of this toxic trinity^{2,3}. The structure shows how EF hijacks one of the key intracellular proteins involved in calcium-triggered signalling pathways, at once disabling the protein and using it to stimulate its own catalytic activity. This activity inhibits the immune response against the bacterium, allowing infection to proceed.

When *B. anthracis* spores are inhaled they lodge in the lungs, where they are ingested by macrophages — immune cells that are the first line of defence against cellular intruders. By mechanisms that are poorly understood, the spores are not destroyed but germinate and multiply within membranous compartments called phagolysosomes in macrophages. The spores then form vegetative cells that eventually kill the macrophages⁴. By that time, however, the macrophages have migrated into the lymph nodes, where their deadly cargo is released. From there, *B. anthracis* moves into the bloodstream, where it multiplies prolifically and secretes anthrax toxin, ultimately killing its host with a form of septic shock.

The LF and PA components of anthrax toxin seem sufficient to cause these later features of an 'inhalation' anthrax infection. So the involvement of EF in such infections has been unclear, although it is known to cause tissue swelling (oedema) in the usually non-lethal, cutaneous form of anthrax. Recent data have indicated, however, that all three toxin components are required for the spores to germinate and survive in macrophages, and then to kill them⁵. Studies of the interaction between anthrax toxin and macrophages have generally looked at how purified toxin proteins enter macrophages from the bloodstream — that is, the mechanisms underlying the late stages of infection. But it is easy to imagine that similar strategies

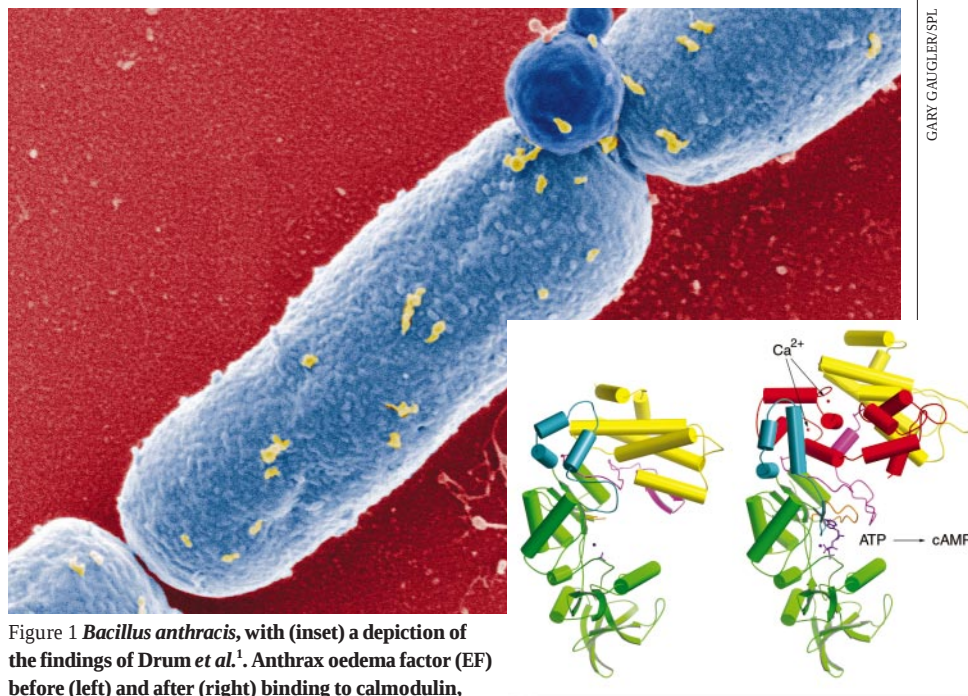


Figure 1 *Bacillus anthracis*, with (inset) a depiction of the findings of Drum *et al.*¹. Anthrax oedema factor (EF) before (left) and after (right) binding to calmodulin, shown in red. The yellow and blue arms of EF grab and twist calmodulin so that it can no longer send signals to other proteins in the host cell. At the same time, binding of calmodulin organizes two loops (orange and purple), creating a binding site for ATP that allows catalytic conversion to cyclic AMP to proceed. (Diagram courtesy of A. Bohm.)

could be used at an earlier stage to transport the toxins into macrophages from within, across the phagolysosome membrane. If so, the transport process would begin with the activation of PA by protein-cleaving enzymes, and the insertion of PA (induced by the acidic conditions in the phagolysosome) into the phagolysosome membrane. EF and LF could then pass through PA into the cell.

EF is an adenylyl cyclase: it catalyses the production of cyclic AMP from ATP, the cell's main energy store. Precisely how increased levels of cyclic AMP mediate the toxic effects of EF is unclear. But, in macrophages, cyclic AMP inhibits the activity of the gene-transcription factor NF- κ B — a protein that would normally stimulate the production of several inflammatory factors that coordinate immune responses. LF, on the other hand, is a surgical protease: it cleaves one specific bond in signalling proteins of the MAPKK family, effectively cutting their lines of communication^{6,7}. Several members of this family cooperate in pathways that converge on the activation of NF- κ B⁸, suggesting how the two toxic enzymes, LF and EF, could

collaborate to defeat the immune system, including promoting macrophage death⁹.

EF is highly efficient at catalysing the production of cyclic AMP, but only when it is bound to the host protein calmodulin. Presumably, such *in situ* activation is important for the bacterium to avoid the effects of its own toxin. But why use calmodulin? First, it is highly abundant, representing up to 1% of total cellular protein. Second, it is the principal mediator of Ca^{2+} signalling in cells. In response to several different extracellular stimuli, Ca^{2+} levels rise subtly but significantly, from resting levels of 10^{-7} M to 6×10^{-6} M. Calmodulin contains two lobes: one has high affinity for Ca^{2+} and binds to two Ca^{2+} ions all the time; the other has lower affinity and binds only when intracellular Ca^{2+} levels rise. Binding of Ca^{2+} to the second lobe brings about a change in structure that exposes hydrophobic surfaces in calmodulin. This allows it to bind to and activate other signalling molecules — usually by wrapping around an autoinhibitory helical structure, or a bundle of helices, in these molecules.

Drum *et al.*'s crystal structure of EF in

complex with calmodulin¹ reveals that the toxin rides roughshod over these subtle molecular signals, and that it binds to calmodulin in a very unusual way. First, EF binds to calmodulin in the 'low calcium' state, even at the very high (millimolar) levels of Ca^{2+} used to grow crystals. It does this by gripping the lower lobe with one arm while another arm grabs the upper lobe and twists it relative to the first (Fig. 1). This generates a contact region of 6,000 Å² between the two proteins, the largest such region observed so far.

However, although the binding constant for the EF–calmodulin complex is good enough to ensure an irreversible complex, it is lower than that expected from the size of the contact region. Presumably, some of the binding energy must be expended in the large conformational changes that are induced in both EF and calmodulin. This molecular 'full nelson' locks calmodulin into a conformation that cannot bind to its normal intracellular targets, even at increased Ca^{2+} levels. This raises the possibility that EF serves to soak up or 'titrate out' calmodulin, preventing it from carrying out its usual activities. These include the activation of enzymes that catalyse the breakdown of cyclic AMP to form ATP — the reverse of the reaction catalysed by adenylyl cyclase.

Meanwhile, the new contacts between EF and calmodulin order and reorganize two loops of EF to allow it to bind to its substrate (ATP; Fig. 1). In an example of convergent evolution, the structure reveals that EF's catalytic machinery has similarities, as well as differences, to that of eukaryotic adenylyl cyclases, although the structure of the protein surrounding and supporting the active site is quite different. This is good news for drug designers, although it remains to be seen whether EF is a useful drug target.

What is left to be done? Much of the work on anthrax toxin has involved the binary combinations of either EF with PA or LF with PA. But it seems increasingly likely that EF and LF work together to alter and disable signalling pathways that are involved in host defence, particularly during the critical early stages of infection. This molecular teamwork will no doubt be the subject of much future research. ■

Robert C. Liddington is in the Program on Cell Adhesion–Extracellular Matrix Biology, The Burnham Institute, 10901 North Torrey Pines Road, La Jolla, California 92037, USA.

e-mail: rliddington@burnham-inst.org

1. Drum, C. L. *et al.* *Nature* **415**, 396–402 (2002).
2. Petosa, C., Collier, R. J., Klimpel, K. R., Leppla, S. H. & Liddington, R. C. *Nature* **385**, 833–838 (1997).
3. Pannifer, A. D. *et al.* *Nature* **414**, 229–233 (2001).
4. Dixon, T. C., Fadl, A. A., Koehler, T. M., Swanson, J. A. & Hanna, P. C. *Cell. Microbiol.* **2**, 453–463 (2000).
5. Guidi-Rontani, C., Levy, M., Ohayon, H. & Mock, M. *Mol. Microbiol.* **42**, 93–98 (2001).
6. Duesbery, N. S. *et al.* *Science* **280**, 734–737 (1998).
7. Vitale, G. *et al.* *Biochem. Biophys. Res. Commun.* **248**, 706–711 (1998).
8. Delgado, M. & Ganea, D. *J. Biol. Chem.* **276**, 369–380 (2001).
9. Kyriakis, J. M. *Nature* **414**, 265–266 (2001).

Extragalactic astronomy

Cosmic censorship

Max Tegmark

Cosmologists believe the Universe will expand forever, but our powers to observe it will not accelerate at the same rate. What does this mean for the future of astronomy?

The night sky seen through a large telescope is adorned with billions of galaxies in a rich variety of shapes and sizes. Yet 100 billion years from now, the sky is likely to be dark and drab, with only a handful of galaxies still visible to us. The discovery in 1998 that the expansion of the Universe is accelerating^{1,2} (rather than slowing down or even reversing) has generally been heralded as good news for our long-term future, because it suggests that the Universe will expand forever, rather than end in a cataclysmic crunch. But it has long been known that such a future is fraught with problems that are likely to make it fairly lonely. Apart from the inevitable death of individual objects, such as our Sun running out of fuel, all but our closest neighbours will disappear forever across a 'cosmic event horizon' from beyond which no light or information will ever reach us. In a paper available as a preprint, and appearing soon in *Physical Review D*, Abraham Loeb³ has quantified this cosmic censorship to a new level of detail, and his results provide sobering news for extragalactic astronomers.

This is how the cosmic censorship works. Suppose for a moment that we lived in a much simpler, non-expanding, infinite Universe of the sort scientists imagined 100 years ago. If such a Universe had been created in a Big Bang 13 billion years ago, then we could never see further than 13 billion light years in any direction, because light from further away would not have had time to reach us. We would effectively be surrounded by a spherical horizon with a radius of 13 billion light years, which would grow linearly with time so that we could see almost any distance if we just waited long enough. The same thing happens in an expanding Universe — there is still a spherical horizon from beyond which light has not had time to reach us. Indeed, the expansion eliminates the question of what the Big Bang looked like: distant objects become progressively dimmer, with the horizon itself infinitely dim and hence black and invisible.

Until recently, there was no reason for astronomers to feel censored: the expansion of the Universe was expected to keep slowing down, so our visible horizon would keep growing without bound, encompassing ever more objects (Fig. 1). But the discovery that the expansion of the Universe is actually speeding up (as a result of mysterious 'dark energy') changes all this, causing the growth



Figure 1 Here today, gone tomorrow: galaxies visible to the Hubble Space Telescope. Does the accelerating expansion of the Universe signal a bleak future for astronomy? Abraham Loeb³ calculates that in 100 billion years most of the currently observable Universe will be invisible to us.

of the horizon to grind to a halt. Because distant galaxies will recede from us at ever increasing rates, they will soon leave the region of spacetime that we will be able to observe. The only galaxies that will stay and keep us company are those to which we are gravitationally bound, such as those in the Virgo cluster.

But galaxies will not simply pop out of sight. Imagine a spiral galaxy spinning as it recedes from us, appearing smaller and dimmer over time. As we see it approach the horizon, it will gradually appear to stop shrinking and spinning, its image forever frozen in time. The frozen image will then slowly become dimmer and fade from view. This means that we will never be able to see most objects develop beyond a certain age. This is analogous to what something looks like when it falls through the event horizon into a black hole, only this time the event horizon is around us, and we are forever trapped inside — doomed to experience a sort of cosmic claustrophobia.

Although not the most urgent problem facing humankind, future astronomers will be severely limited in what they can see and do, as Loeb shows³. For instance, if we

ROBERT WILLIAMS/HUBBLE DEEP FIELD TEAM/NASA

detect intelligent beings in a galaxy with a redshift (an astronomical proxy for distance) exceeding 1.8 and send them a message, it would never arrive. If we point a telescope at a distant galaxy whose current redshift is 5 (corresponding to a time when the Universe was less than 2 billion years old), we will never see it get older than about 6 billion years. This means that arguments relating the properties of young sources at high redshift to those of their present-day counterparts can never be directly confirmed. This galaxy's redshift evolves dramatically. After decreasing in the past, reaching its lowest value around now, it will grow to 6 as the Universe doubles its age, hit 27 after another age-doubling, and gradually approach infinity as the galaxy fades from sight. To put all this in context, the Sun will die about 5 billion years from now, and most galaxies will have dimmed dramatically in 100 billion years, as even the most long-lived stars exhaust their fuel supply.

There is still a glimmer of hope for cosmic

claustraphobics. Although the evidence for dark energy is mounting — its current strength (energy density) has now been fairly accurately measured^{4,5} — its physical nature is still a mystery. The data so far are consistent with the density of dark energy staying constant over time, as Loeb assumes in his paper, but the possibility remains that it might decay quickly enough for the horizon to grow indefinitely. Nonetheless, until dark energy is better understood, astronomers are advised to play it safe and enjoy the heavens while they can. To paraphrase Horace: seize the night!

Max Tegmark is in the Department of Physics and Astronomy, University of Pennsylvania, Philadelphia, Pennsylvania 19104-6396, USA.
e-mail: max@physics.upenn.edu

1. Perlmutter, S. *et al.* *Nature* **391**, 51–54 (1998).
2. Riess, A. G. *et al.* *Astron. J.* **116**, 1009 (1998).
3. Loeb, A. *Phys. Rev. D* (in the press); <http://xxx.lanl.gov/abs/astro-ph/0107568>
4. Wang, X., Tegmark, M. & Zaldarriaga, M. <http://xxx.lanl.gov/abs/astro-ph/0105091>
5. Efstathiou, G. *et al.* <http://xxx.lanl.gov/abs/astro-ph/0109152>

Photosynthesis

Evolutionary options

J. A. Raven

The discovery in tobacco and celery of components of the C₄ type of photosynthesis presents a host of intriguing questions about when and how this pathway evolved in higher plants.

Higher plants can be divided into two groups — C₃ and C₄ — according to which biochemical route they use to produce their food from water and carbon dioxide (the numbers simply refer to carbon atoms in the initial photosynthetic product). Over 85% of higher plants use the C₃ route¹; the C₄ pathway, which has evolved at least 30 times, seems to be a response to low levels of CO₂ in the atmosphere¹. These multiple origins suggest that the enzymatic and structural prerequisites of C₄ metabolism were widespread, at least in the 18 extant plant families with C₄ members. On page 451 of this issue², Hibberd and Quick describe how they have identified cells in C₃ plants that use part of the C₄ pathway. In doing so, they provide clues to the structure, metabolism and function of C₄ precursor structures.

The current increase in atmospheric CO₂ has focused attention on how vegetation responds to CO₂. In C₃ plants, the short-term rate of carbon assimilation responds directly to the external CO₂ concentration. In the other 15% of plants, the rate of photosynthesis does not increase in response to rising levels of atmospheric CO₂ because the plants assimilate carbon in a way that accumulates CO₂ and minimizes its loss. This is the C₄ photosynthetic pathway, or a variation on C₄ known as crassulacean acid metabolism,

which are used by 7,500 and 30,000 species, respectively. Species of both C₃ and C₄ plants are of great agricultural importance.

What are the main features of C₄ photosynthesis? And so what enzymatic and structural components would have been required in the ancestral plants? The enzyme ribulose biphosphate carboxylase-oxygenase (Rubisco for short) is at the core of CO₂ assimilation in all plants. In C₃ land plants, atmospheric CO₂ is supplied to Rubisco in leaf (mesophyll) cells, mainly by diffusion of gas into the leaf and then in solution through a micrometre or two of water. The initial, three-carbon product then feeds into a biochemical cycle to produce sugars.

In typical C₄ plants^{1,3}, Rubisco tends to occur not in the mesophyll but in a layer of tissue surrounding the vascular bundle — this contains the plant's water-conducting (xylem) and food-conducting vessels, and the surrounding tissue is composed of bundle sheath cells. There are no gas spaces between these cells, and CO₂ diffusion to and from them is very restricted. In C₄ plants, atmospheric CO₂ is initially fixed in mesophyll cells by the enzyme phosphoenolpyruvate carboxylase (PEPC), producing the four-carbon acids malate or aspartate. These products pass through tubes between cells to the bundle sheath cells, where CO₂ is released

in the presence of Rubisco by one of three decarboxylase enzymes. The remaining three-carbon compound is recycled to the mesophyll cells and is used to fix more CO₂ via PEPC. Most of the CO₂ released into the bundle sheath cells is fixed by Rubisco at concentrations several times greater than that in air, so energy-consuming photorespiration is inhibited at the energetic expense of pumping CO₂ into the bundle sheath cells.

Hibberd and Quick² investigated the metabolism of green cells in vascular bundles in the stems and leaf stalks of two C₃ plants, *Nicotiana tabacum* (tobacco) and *Apium graveolens* (celery). These were tested as possible precursors of the C₄ bundle sheath. Hibberd and Quick labelled CO₂ and malate, both of which occur naturally in xylem sap, with ¹⁴C. They show that ¹⁴C from both of these sources is assimilated photosynthetically into sugars by the green cells using Rubisco. The authors also investigated the activities of decarboxylases, which could act on C₄ dicarboxylic acids in the green cells of vascular bundles in tobacco, and found higher levels of all three enzymes than in leaf cells. These findings are depicted in Fig. 3 of the paper (page 452).

Why do C₃ plants have this C₄ capacity? First, why do they photosynthetically convert xylem CO₂ into sugars, using vascular bundle cells? The short answer is that it is a way of conserving carbon. The respiration of both roots and soil organisms produces CO₂, which diffuses into the xylem sap. This sap rises. When it reaches the parts of the plant that are above ground, there is a CO₂ concentration gradient between the sap and the atmosphere which results in the loss of CO₂ by diffusion. The CO₂ conductance in this radial pathway is low, especially in the vascular bundles where gas spaces are absent, favouring fixation by Rubisco over loss to the atmosphere.

This mechanism of trapping CO₂ works only in the light, perhaps because CO₂ flux up the xylem (concentration of CO₂ times speed of movement of the xylem stream) is greater in the light, when most water loss from leaves occurs in C₃ plants. So this is when most CO₂ could be lost from the xylem. The CO₂ flux from the roots can be 1–7% of that from the atmosphere into leaves⁴, meaning that the recovery of xylem CO₂ is significant in the whole-plant carbon balance⁵.

Explanation is also required for the decarboxylation of root-produced malate in bundle sheath cells. This is less readily rationalized, because Hibberd and Quick² point out that this malate helps to maintain the acid–base balance in shoots. It would be useful to carry out experiments of the type they performed, but using plants with different nitrogen sources that have different effects on the acid–base balance in whole plants.

If the green cells in vascular bundles are

precursors of the bundle-sheath components of C_4 photosynthesis, this raises some evolutionary questions. When did these cells first occur, and so when might this route to C_4 metabolism have first been available? Decreasing CO_2 levels over the past 20 million years or so, in the Upper Tertiary and Quaternary periods, are known to have been an important factor in the evolution of C_4 metabolism in one of the main groups of higher plants, the flowering plants. But there were also low atmospheric CO_2 levels in the Upper Palaeozoic some 300 million years ago, over 100 million years after the first known higher plants⁶. Oxygen levels in the Upper Palaeozoic atmosphere were higher than today⁷. Combined with low CO_2 levels, this would have decreased the rate of C_3 photosynthesis as a result of the kinetic properties of Rubisco⁸, and favoured the evolution of C_4 metabolism. But C_3 metabolism was the main — perhaps the only⁹ — mechanism of photosynthesis in the Upper Palaeozoic.

The Upper Palaeozoic forerunners of two of the other major groups of higher plants, the early ferns and conifers, probably did not use the C_4 pathway. Might that have been because the anatomy and physiology of their primary vascular tissue were inappropriate for the photosynthetic mechanism identified by Hibberd and Quick? That is one possibility. But despite the abundant fossil evidence of vascular tissue in Upper Palaeozoic plants, it is not clear whether the associated cells could photosynthesize. So another possibility is that the default condition for the cells of all higher plants, when exposed to the light, was the development of photosynthetic capacity; suppressing that capacity was an evolutionarily derived state. On these grounds, photosynthesis in vascular tissue would have been at least as common in the past as it is now, when chlorophyll-containing cells are associated with vascular tissue in 30 families.

Hibberd and Quick² have not said the last word on the evolution of C_4 photosynthesis. But their work makes a significant contribution, not least in pointing to several new lines of investigation.

J. A. Raven is in the Division of Environmental and Applied Biology, School of Life Sciences, Biological Sciences Institute, University of Dundee, Millers Wynd, Dundee DD1 4HN, UK.
e-mail: j.a.raven@dundee.ac.uk

1. Sage, R. F. *Plant Biol.* **3**, 202–213 (2001).
2. Hibberd, J. M. & Quick, W. P. *Nature* **415**, 451–454 (2002).
3. Voznesenskaya, E. V. *et al.* *Nature* **414**, 543–546 (2001).
4. Levy, P. E. *et al.* *Tree Physiol.* **19**, 53–58 (2001).
5. Levy, P. E. & Jarvis, P. G. *Funct. Ecol.* **12**, 107–116 (1998).
6. Berner, R. A. & Kothavala, Z. *Am. J. Sci.* **301**, 182–204 (2001).
7. Berner, R. A. *Geochim. Cosmochim. Acta* **65**, 685–694 (2001).
8. Beerling, D. J. & Berner, R. A. *Proc. Natl Acad. Sci. USA* **97**, 12428–12432 (2000).
9. Wright, V. P. & Vanstone, S. D. *J. Geol. Soc. Lond.* **148**, 945–947 (1991).

Huntington's disease

Accomplices to neuronal death

Mark P. Mattson

The shaky movements of people with Huntington's disease are caused by death of a specific group of nerve cells. One culprit — the mutant huntingtin protein — is well known. Another has now been discovered.

When you lift your coffee cup or turn the pages of this journal, you are carrying out a 'voluntary' action. Such movements are produced in response to complicated sequences of nerve cells, and are coordinated by the striatum region of the brain. The job of this region is to inhibit particular neurons in the brain and spinal cord that are not required for a given action. So damage to the striatum can result in uncontrolled body movements — a scenario that occurs in people with Huntington's disease. This inherited, age-related disease is caused by the abnormal lengthening of the repeated sequence CAGCAGCAG... in the gene encoding the huntingtin protein, resulting in long stretches of the amino acid glutamine within the protein. Expression of mutant huntingtin in mice leads to the death of striatal neurons and a neurological disorder that is similar to Huntington's disease¹. Although the exact details are unclear, the molecular and cellular mechanisms by which mutant huntingtin kills striatal neurons are beginning to be unravelled. Writing in *Nature Cell Biology*, Gervais and colleagues² add more pieces to the puzzle.

It was already known that mutant huntingtin somehow triggers apoptosis — a process by which cells effectively commit suicide, systematically chewing up their DNA and degrading their proteins and organelles. In fact, apoptosis is increasingly being implicated in neurodegenerative disorders, from Alzheimer's and Parkinson's diseases to strokes³. The abnormality that triggers apoptosis may be different in each disorder; for example, Alzheimer's disease is characterized by the abnormal accumulation of amyloid- β peptide, and neuronal deaths in strokes are triggered by a shortage of oxygen supply to the brain. But in each case apoptosis is carried out by cascades of protein-cleaving enzymes called caspases. In particular, caspase-3 and caspase-8 have

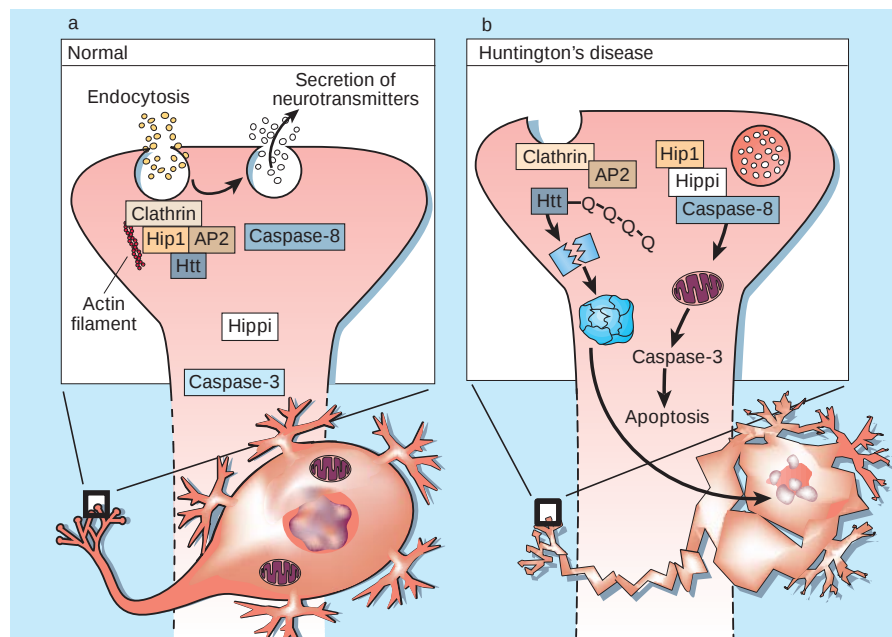


Figure 1 Possible functions of the normal and mutant huntingtin protein. **a**, In normal nerve cells, huntingtin (Htt) can form a complex with the proteins Hip1, clathrin and AP2. Clathrin and AP2 (and so perhaps Htt) are involved in endocytosis — the internalization of sections of the plasma membrane — which is crucial in allowing neurons to communicate by secreting neurotransmitters. **b**, In patients with Huntington's disease, the Htt protein has an abnormally long tract of glutamine (Q) amino acids. This mutant protein may lead to abnormal endocytosis and secretion in neurons. In addition, it causes striatal neurons to die by the process of apoptosis. Gervais *et al.*² show that the long tract of glutamines weakens Htt's interaction with Hip1, which is then free to bind the protein Hippi, and so to activate apoptosis through caspase-8 and caspase-3. Caspase-3 also cleaves Htt, producing fragments that clump together and form inclusions in the neuron and its nucleus.



100 YEARS AGO

We are glad to learn that the gliding experiments with which Lilienthal and Pilcher sought to investigate the balance and stability of machines supported by *aéroplanes* and *aërocurves* have not been discontinued since the death of these two investigators. A great deal of valuable work has already been done in America by Mr. Octave Chanute, and in conjunction with him by Mr. Herring, both of whom have attained results in advance of those previously achieved, by the use of machines provided with movable wings. Still more recently, i.e. from October 1900 onwards, two other workers have attacked the problem, namely, Mr. Wilbur Wright and Mr. Orville Wright, of Dayton, Ohio. Mr. Wilbur Wright adopts a two-surfaced machine and assumes a horizontal position when gliding, with the view of diminishing head resistance. He has successfully worked with a surface area of double that used by previous experimenters, and has on several occasions extricated himself from the dangerous position in which Lilienthal and other observers have found themselves when suddenly brought to rest in a high wind.

From *Nature* 23 January 1902.

50 YEARS AGO

Is there an æther? In his recent communication Prof. Dirac has said that in his new formulation of electrodynamics a preferred motion exists at each point of space. A preferred motion is also given at each point of space by cosmological observations (apparent isotropy of distant red-shift effects). It is of interest whether any local physical effects are associated with this cosmologically preferred state of motion, and the analogous question can be asked with respect to Dirac's formulation. We have argued that the first question may be answered in terms of the theory of continual creation (the steady-state theory of the universe), where the cosmologically preferred motion is identified with the velocity of newly created particles. Dirac answers the second question by saying that a small charge placed in a vacuum would possess the particular velocity associated with the potentials in his theory. The concept of introducing a charge into a vacuum, without fields destroying the vacuum character of the region, is given physical reality in a theory of continual creation.

H. Bondi & T. Gold

From *Nature* 26 January 1952.

been implicated in neuronal death in such diseases³.

Caspases are also presumed to be involved in Huntington's disease. They are activated in striatal neurons in Huntington's patients, and in cell-culture and animal models of the disease⁴. Moreover, drugs that inhibit caspases can protect striatal neurons from death in such model systems. But the specific mechanism by which mutant huntingtin triggers the apoptotic caspase cascade has been a mystery, with insufficient evidence available to indict any particular suspect. Previous studies⁴ had found that mutant huntingtin leads to the activation of caspase-8. Other investigators⁵ have shown that normal huntingtin interacts with a protein known as Hip1. But it was unclear whether and how that interaction is related to Huntington's disease.

Gervais and colleagues² now provide an answer. Their studies of cultured cells show that lengthening the tract of glutamine amino acids in huntingtin weakens this protein's interaction with Hip1 to the extent that it is released. Hip1 then interacts with a newly identified protein, Hippi, that appears to be essential for forming a 'death-effector complex' involving caspase-8, thereby setting off the apoptotic cascade. In this cascade, activation of caspase-8 can trigger alterations in mitochondria—the cellular powerhouses—which then release the cytochrome *c* protein; this in turn forms a protein complex called the 'apoptosome', which activates caspase-3. Caspase-8 may also bypass mitochondria and activate the apoptosome directly. Interestingly, caspase-3 cleaves huntingtin, releasing a fragment that can self-aggregate and form inclusions in the cell and nucleus. It is unclear whether these inclusions are required for neuronal death in Huntington's disease, because caspase-3 can kill neurons independently.

Gervais *et al.*'s results² show that Hippi is critical for the death-promoting effects of mutant huntingtin in cultured cells. But considerable further work will be required to establish whether this occurs in Huntington's patients. We need to know in what cells and at what levels Hippi is expressed normally in the human brain, and whether its expression or subcellular localization is changed in Huntington's patients in a way that is related to the selective vulnerability of striatal neurons. Another question is whether there is any relationship between huntingtin, Hip1, Hippi and the perturbed cellular energy metabolism that is a major feature of Huntington's disease⁶.

Moreover, inherited expansions of tracts of 'trinucleotide repeats' (such as CAG in Huntington's disease) occur in other genes, and in several cases these expansions also cause neurological disorders characterized by the dysfunction and death of neurons that control body movements. Two examples are spinocerebellar ataxia and Kennedy's disease⁷. It remains to be seen whether Hip1

and Hippi have similar roles in triggering apoptosis in these disorders. Finally, alterations in mitochondria are central to the development of Huntington's disease and perhaps to neuronal death. So a further question is whether the Hippi-caspase pathway lies upstream of these mitochondrial changes. If so, how does this pathway interact with other proteins, such as Par4 and members of the Bcl2 protein family^{3,8}, that control apoptosis at a pre-mitochondrial step?

Beyond their potential implications for preventing and treating Huntington's disease, studies of the pathogenic effects of mutant huntingtin are also revealing the normal function of this protein in neurons (Fig. 1). Several findings are consistent with huntingtin being involved in regulating 'synaptic plasticity'—the ability of neurons to change the strength of their connections (synapses) with other neurons. When neurons communicate, small membrane-clad sacks filled with neurotransmitters in the signal-transmitting (presynaptic) neuron fuse with the neuron's plasma membrane. The neurotransmitters are thereby released and diffuse away, to be picked up by receptor molecules on the extensions (dendrites) of a receiving (postsynaptic) neuron. Several membrane-transport processes are involved, including vesicle fusion with the plasma membrane and the internalization (by 'endocytosis') of excess plasma membrane. The receptors on the receiving neuron can also be internalized by endocytosis and recycled.

Proteins called clathrin and adaptor protein-2 (AP2) are needed for endocytosis, and clathrin-rich endocytotic zones are present in presynaptic nerve terminals, where they may link membrane components with the actin-based cellular 'skeleton'⁹. Hip1 interacts with both AP2 and clathrin¹⁰, and it seems that huntingtin is associated with clathrin-coated membranes along dendrites¹¹. The implication is that normal huntingtin has both presynaptic and postsynaptic functions, possibly in processes such as neurotransmitter release and receptor recycling—all of which may contribute to synaptic plasticity.

Gervais *et al.*'s results² suggest that mutations in huntingtin might, by reducing its interaction with Hip1, alter endocytotic or secretory processes, leading to synaptic dysfunction and the activation of caspase cascades specifically at synapses. Indeed, apoptotic biochemical cascades involving caspases can be triggered in specific regions of neurons, and may then propagate the cell-death signal to the nucleus¹². The local activation of caspases may also affect synaptic plasticity and the structural remodelling of synapses¹³. So, as details of the cellular and molecular alterations that sentence neurons to death in trinucleotide-repeat disorders are revealed, so too are the intricate mechanisms that allow neuronal circuits to carry

out their functions and adapt to the environmental demands of ageing, such as impaired energy metabolism.

Mark P. Mattson is at the National Institute on Aging, Gerontology Research Center, 5600 Nathan Shock Drive, Baltimore, Maryland 21224, USA. e-mail: mattsonm@grc.nia.nih.gov

1. Hackam, A. S. *et al.* *Phil. Trans. R. Soc. Lond. B Biol. Sci.* **354**, 1047–1055 (1999).
2. Gervais, F. G. *et al.* *Nature Cell Biol.* 14 January 2002. (DOI 10.1038/ncb735).

3. Mattson, M. P. *Nature Rev. Mol. Cell Biol.* **1**, 120–129 (2000).
4. Sanchez, I. *et al.* *Neuron* **22**, 623–633 (1999).
5. Wanker, E. E. *et al.* *Hum. Mol. Genet.* **6**, 487–495 (1997).
6. Beal, M. F. *Trends Neurosci.* **23**, 298–304 (2000).
7. Evert, B. O., Wullner, U. & Klockgether, T. *Cell Tissue Res.* **301**, 189–204 (2000).
8. Duan, W., Guo, Z. & Mattson, M. P. *Exp. Neurol.* **165**, 1–11 (2000).
9. Wenk, M. R. *et al.* *Neuron* **32**, 79–88 (2001).
10. Metzler, M. *et al.* *J. Biol. Chem.* **276**, 39271–39276 (2001).
11. Velier, J. *et al.* *Exp. Neurol.* **152**, 34–40 (1998).
12. Mattson, M. P., Keller, J. N. & Begley, J. G. *Exp. Neurol.* **153**, 35–48 (1998).
13. Lu, C., Fu, W. & Mattson, M. P. *Neuromol. Med.* **1**, 55–65 (2002).

High-temperature superconductivity

Quantum salad dressing

Jan Zaanen

The mystery of how electrons in a high-temperature superconductor flow without resistance grows deeper. New pictures at the atomic scale reveal two electronic phases that — like oil and vinegar — do not easily mix.

According to quantum mechanics, when space is the same everywhere, a particle should be everywhere at the same time. In low-temperature metals and superconductors, this principle rules the waves of the quantum fluids formed by the electrons. Under such conditions, electronic fluids seem remarkably featureless and insensitive to imperfections in the material. For instance, in a conventional superconductor the structure of the crystal lattice must be severely messed up to get a noticeable response from the superconducting electrons. But the high-temperature superconductivity found in certain copper oxide compounds is different. Quantum physics is still in charge but the underlying quantum fluids do not behave like their low-temperature counterparts. On page 412 of this issue, Lang *et al.*¹ take a direct look at this electronic fluid at the nanometre scale, using scanning tunnelling microscopy (STM)². Instead of the implacable perfection of the conventional superconductor, they find it to be rather messy.

What does the fluid look like? Imagine pouring salad dressing (a mixture of oil and vinegar) on to a plate covered with spots of fat. The result is droplets of vinegar immersed in an oily matrix. If the vinegar were superconductor and the oily matrix were a mystery state of electronic matter called the pseudogap phase, the result would be similar to what Lang *et al.* saw. (The fatty spots represent imperfections in the crystal lattice of the superconductor.) Until now, all we've been able to see at the macroscopic scale is the superconducting droplets merging together, as if the plate's contents as a whole turn into pure vinegar. This merging occurs because of a quantum effect called Josephson coupling. The nanoscale images obtained by Lang *et al.* help us probe beyond the macroscopic surface of this electron system.

Understanding the quantum behaviour of a single electron is hard enough, but there

are approximately 10^{23} strongly interacting electrons in every gram of earthly matter. Fortunately the basic rules simplify when the system becomes big. The secret of conventional quantum fluids is that the particles seem to forget that they interact when the system size is scaled up. But they keep their quantum-mechanical habit of spreading out across all the space, and their quantum-statistical nature: fermions (electrons) turn into a Fermi gas (normal metals), and bosons (pairs of electrons) turn into a Bose–Einstein condensate (superconductors).

This is what lies behind the implacable perfection of conventional quantum fluids: the system of particles inherits its basic properties directly from its constituents. But condensed-matter physics is increasingly concerned with quantum systems that do not conform to this description — such as the electron system of the high-temperature superconductors. Despite unprecedented research activity (approximately 100,000 papers since its discovery in 1987), high-temperature superconductivity seems as mysterious as ever. Part of its continuing fascination for physicists is its potential to reveal fundamental insights into the collective behaviour of quantum particles^{3,4}.

I've already given away the punch-line of Lang and colleagues' result¹: the electron system of their superconductor looks like salad dressing. But there is deeper meaning to this kitchen-table metaphor — the STM pictures are messy because the superconducting electrons do not forget their interactions when they turn into a quantum fluid. Instead, the electronic matter is still strongly interacting, causing it to act in a highly collective way. Such collective behaviour carries its own logic that supersedes even the differences between classical and quantum physics, allowing me to use everyday phenomena to describe what is happening.

To explain further, salad dressing starts

out as protons, neutrons and electrons. If these were non-interacting, salad dressing would remain featureless. Instead, these entities bind into lipid, acetic acid and water molecules, which in turn form two strongly interacting, immiscible fluids: vinegar and oil. Something similar happens in the high-temperature superconductors. In the beginning there are electrons and some crystal-lattice vibrations. These are strongly interacting and the electrons quickly lose their identity and morph into a collective phase with its own emergent properties. Apparently, there are two of these phases — the 'superconducting' and the 'pseudogap' phases — but they do not easily mix. This highly collective electron matter reacts strongly to imperfections in the superconductor (probably missing oxygen atoms in the crystal lattice), causing the droplet picture.

The superconducting phase shares at least some properties with a conventional superconductor, but the pseudogap phase is still a mystery. There are several reasons why it could exist. Initially the pseudogap phase was thought to correspond to a high-temperature precursor of the superconducting state, in which the electrons formed pairs that did not undergo Bose–Einstein condensation until the temperatures dropped low enough². More recently it was thought to be a novel quantum state of matter characterized by an exotic form of order that is difficult to observe experimentally (hidden order)⁴. Theorists came up with several ingenious proposals regarding the nature of this order, such as the flux and *d*-density wave orders^{5,6}. Lang *et al.*'s observation of a two-phase mixture at low temperatures makes the high-temperature-precursor explanation less likely, as the pseudogap phase can exist simultaneously with the superconducting state. But it does give strong support to the notion that the pseudogap phase is a separate state of matter.

Lang *et al.* also discovered a very strange property of the pseudogap phase. To investigate the differences between the two phases they introduced nickel impurities into some of their samples. These impurities produce STM fingerprints in the superconducting phase known as 'impurity resonances'. According to theory, the authors should have found a similar, but modified, pattern of impurity resonances in the pseudogap regions. Instead, their measurements show that the resonances disappear completely. This is a serious challenge for existing theories, which are invariably based on the assumption that both phases are very similar on atomic scales, with differences emerging only on larger scales. So if the impurity resonances look drastically different in the pseudogap and superconducting phase, the phases must be dissimilar at the atomic scale. Again the salad-dressing metaphor is helpful — it seems that only a few

electrons are needed to form distinguishable oil (pseudogap) and vinegar (superconductivity) 'molecules'.

What good might come of all this? Studies like these are revealing amazing diversity in the behaviour of these mystery electron systems at the nanoscale. Indeed, much more appears to be going on than in more conventional systems, such as gold, silicon and carbon nanotubes, which form the building blocks of current nanotechnology. Perhaps one day the rich electronic nano-life of high-temperature superconductors will

also be tamed and exploited. We are not just solving a mystery, but witnessing the birth of a new frontier of nanoscience. ■

Jan Zaanen is at the Instituut Lorentz for Theoretical Physics, Leiden University, POB 9504, 2300 RA Leiden, The Netherlands.
e-mail: jan@lorentz.leidenuniv.nl

1. Lang, K. M. *et al.* *Nature* **415**, 412–416 (2002).
2. Yazdani, A. & Lieber, C. M. *Nature* **401**, 227–230 (1999).
3. Buchanan, M. *Nature* **409**, 8–11 (2001).
4. Zaanen, J. <http://arXiv.org/abs/cond-mat/0103255>
5. Wen, X.-G. & Lee, P. A. *Phys. Rev. Lett.* **76**, 503–506 (1996).
6. Chakravarty, S., Laughlin, R. B., Morr, D. K. & Nayak, C. *Phys. Rev. B* **63**, 094503–094513 (2001).

Nitrogen cycle

Natural organic tendency

Nico van Breemen

Dissolved nitrogen is usually thought to be transported from land to sea in inorganic form. But the predominance of organic nitrogen in streams in temperate South America suggests that this view needs a rethink.

The Haber–Bosch process for transforming atmospheric N_2 into ammonia, and so creating nitrogen fertilizer, was invented almost a century ago. Nitrogen is the main nutrient that plants take up from soils, so the Haber–Bosch process was in large part responsible for the vast increase in food production that helped to support a booming human population during the twentieth century. But it has also dramatically altered the global nitrogen cycle. The latest research attesting to that point, by Perakis and Hedin¹, appears on page 416 of this issue.

In ecosystems that have been subject to intensive human activity, such as the temperate Northern Hemisphere, nitrogen that leaches from ecosystems is largely in inorganic form (nitrate). But Perakis and Hedin find that in temperate parts of Chile and Argentina, which have remained free of such human intervention, the nitrogen is mainly organic in form. It seems, then, that our tendency to carry out research in northern regions has given us a misleading picture.

Over the past hundred years, human activity has doubled nitrogen input into the global terrestrial nitrogen cycle. In consequence, in heavily industrialized and densely populated areas we have polluted the atmosphere with ammonia and nitrogen oxides; increased atmospheric deposition of nitrogen; acidified soils, streams and lakes; decreased biodiversity by affecting plants adapted to the efficient use of nitrogen; and affected coastal marine fisheries². These consequences have prompted research into many aspects of the nitrogen cycle, and the results have provided fairly detailed nitrogen budgets for intensively studied areas in the Northern Hemisphere³. Such budgets differ greatly from those that might apply in the areas studied by Perakis and Hedin¹,

quantitatively as well as qualitatively (Fig. 1).

A central feature of nitrogen budgets is the formation of nitrate by a series of biological processes, starting with the decomposition of organic nitrogen in plant litter to ammonium. Ammonium is partly taken up again by plants and partly oxidized by bacteria to nitrate. Nitrate that is not taken up by plants is either leached into rivers or reduced by denitrifying bacteria, mainly to atmospheric N_2 . Denitrification closes the

nitrogen cycle, which started with biological or industrial fixation of N_2 to forms that are available to plants and microbes (most of which cannot use N_2).

The nitrogen cycle is intimately linked to the carbon cycle. Biogeochemical models that describe transformations of either cycle on regional to global scales feature inorganic nitrogen as the dominant feature of nitrogen cycling. These models are commonly used to describe the effects of global change^{4–6}, including those that bear on the question of how much carbon is sequestered on land in response to increased CO_2 in the atmosphere⁷.

Perakis and Hedin¹ repeatedly sampled and analysed stream water in parts of Chile and Argentina (see the map on page 417). They selected one hundred streams, in unpolluted and often remote forested watersheds, looking at ecosystems at different stages of development and subject to a broad range of climatic and geological conditions. They found much more dissolved organic nitrogen than nitrate nitrogen in their water samples, in contrast to a dominance of nitrate in the streams of northern temperate forests (Fig. 1).

It is unlikely that differences in plants and microbes between the hemispheres cause these differences, so the observations call into question several standard concepts about the nitrogen cycle in undisturbed forest ecosystems. Is the dominance of dissolved nitrate, even in moderately polluted northern temperate forests, caused by increased atmospheric deposition there and logging? If it is, this effect has been overlooked by soil

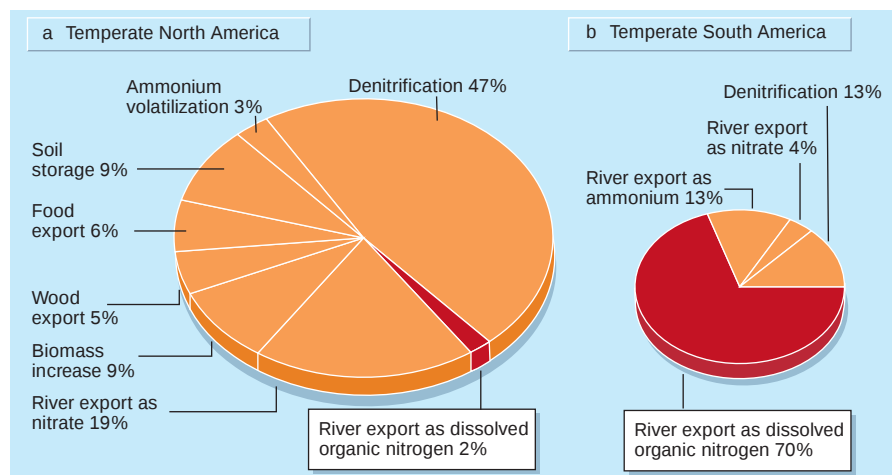


Figure 1 Forms of nitrogen storage in, or export from, two ecosystems in North and South America. a, Data for a largely forested region in the northeastern United States³. Here, nitrogen is found to be exported primarily in inorganic form. b, The pristine forest area in Argentina and Chile studied by Perakis and Hedin¹. Most notably, most of the nitrogen exported in streams and rivers here is in organic, rather than inorganic, form. The two regions have about the same annual precipitation of 1,100 mm. But the total 'sinks' in a are almost ten times those in b (about $36.4 \text{ kg ha}^{-1} \text{ yr}^{-1}$ compared with $0.4 \text{ kg ha}^{-1} \text{ yr}^{-1}$). The difference is due to a much greater input of nitrogen through human agency in the United States. As well as nitrogen export in water, and denitrification, sinks in the forested northeastern United States include export in agricultural and forestry products (food and wood); volatilization of ammonia from manure and fertilizers; storage in growing forests (biomass increase); and accumulation in soils, mainly in forests and in suburban land. The budget in b was estimated from the new data¹, with the assumptions that the nitrogen in soils and biomass is in steady state, and that the ratio of denitrification to nitrate export is the same in both regions.

scientists and ecologists, who largely started to study forest biogeochemistry in earnest only after the onset of widespread extra nitrogen deposition. If the dominance of nitrate is indeed an artefact of human origin, what has triggered the shift from organic to inorganic dissolved nitrogen? And how should we deal with that shift in our global biogeochemical models?

An older and more influential notion is that plants take up nitrogen in inorganic form. This has been a central dogma in plant nutrition for over 150 years⁸, and it is no doubt valid in fertilized farmlands. But the generalization of this concept to natural terrestrial ecosystems in most biogeochemical models is brought into doubt by Perakis and Hedin's findings¹. Indeed, direct uptake of organic nitrogen in different plant groups has been shown to occur in nutrient-poor ecosystems in the far north⁹. Uptake of organic nitrogen may well explain how plants from nutrient-poor ecosystems cope with a preponderance of organic, rather than inorganic, nitrogen in dissolved form in nature.

But first, there is the underlying question of whether nitrogen cycling in northern temperate forests was dominated by dissolved organic nitrogen before the advent of industrial fertilizer. That will be a tough one to answer. All in all, however, Perakis and Hedin's results¹ leave little doubt that some standard thinking about how nature deals with nitrogen in soils and waters needs to be re-evaluated. ■

Nico van Breemen is at the Laboratory of Soil Science and Geology, Department of Environmental Sciences, Wageningen University and Research Centre, POB 37, 6700 Wageningen, The Netherlands.

e-mail: nico.vanbreemen@bodeco.beng.wau.nl

1. Perakis, S. S. & Hedin, L. O. *Nature* **415**, 416–419 (2002).
2. Vitousek, P. M. *et al. Ecol. Appl.* **7**, 737–750 (1997).
3. Van Breemen, N. *et al. Biogeochemistry* (in the press).
4. Herbert, D. A., Rastetter, E. B., Shaver, G. R. & Ågren, G. I. *Ecosystems* **2**, 367–382 (1999).
5. McGuire, A. D. *et al. Glob. Biogeochem. Cycles* **11**, 173–189 (1997).
6. Schimel, D. S., Braswell, B. H. & Parton, W. J. *Proc. Natl Acad. Sci. USA* **94**, 8280–8283 (1997).
7. Fan, S. *et al. Science* **282**, 442–446 (1998).
8. Von Liebig, J. *Chemical Letters*, XV, 2nd corrected edn (Taylor & Walton, London, 1844).
9. Näsholm, T. *et al. Nature* **392**, 914–916 (1998).

Cell biology

Protein sweetener

Armando J. Parodi

After years of failure, a protein involved in flipping certain oligosaccharide molecules into the endoplasmic reticulum — one of the cellular sites where proteins are modified with sugars — has been tracked down.

There are some well-known biological processes that repeatedly resist attempts to characterize them thoroughly. When they finally succumb, it often turns out that an early observation held the key to the solution. The paper by Helenius and colleagues¹ on page 447 of this issue will surely be the key to one such process — how lipid-linked sugar molecules are flipped from the outside to the inside of a cellular compartment known as the endoplasmic reticulum (ER). Such 'oligosaccharide flipping' is needed so that certain newly synthesized proteins can be appropriately tagged with sugar groups (glycosylated). Although this glycosylation initially occurs in the ER, the sugar chains are synthesized on the outside of this compartment, and so need to be transported into it. Helenius *et al.* provide indirect but compelling evidence that a protein called Rft1 is involved.

This protein-tagging process is called 'N-glycosylation', because sugars are covalently attached to the nitrogen (N) of the amido group in the side chain of asparagine amino acids in target proteins. This happens specifically for proteins that are being moved through the compartments of the cellular secretory pathway, from the ER to the Golgi and onwards. In the ER, a preassembled

oligosaccharide containing 14 monosaccharides (three glucoses, nine mannoses and two N-acetylglucosamines) is transferred *en bloc* from a particular lipid (dolichylpyrophosphate, a polyprenol derivative) to protein chains that have not yet folded into their correct shape. As the glycosylated proteins pass through the secretory pathway, their oligosaccharide attachments are substantially modified by the removal and addition of single sugar units (monosaccharides).

In the Golgi, this oligosaccharide remodelling varies greatly from one cell type to another. Moreover, N-linked oligosaccharides on the same or different proteins can also be differently processed within the Golgi of the same cell, yielding mature 'glycoproteins' with sugar moieties of widely differing structures. This diversity reflects how protein-linked oligosaccharides function in mature glycoproteins — mainly in recognition phenomena. For instance, the diversion of lysosomal enzymes from the Golgi to lysosomes depends on specific mannose-6-phosphate groups being on the surface of these glycoproteins.

In the ER, however, both the oligosaccharide transferred from the lipid derivative and the processing reactions that occur vary only

slightly among different cell types. This is because the correct folding of glycoproteins, a process that is common to all such molecules and which occurs in the ER, depends on the presence and structure of oligosaccharides — if proteins are not folded accurately they are retained in the ER and later degraded².

The production of preassembled oligosaccharides presents some interesting topological problems. First, dolichylpyrophosphate-linked Man₃GlcNAc₂ is formed by the successive addition of N-acetylglucosamine (GlcNAc)-phosphate, N-acetylglucosamine and mannose (Man) units onto dolichylmonophosphate, found in the outer leaflet of the ER membrane. This oligosaccharide is then somehow flipped to the inside of the ER, where four more mannose and three glucose units are added. For the first part of this process, the mannose units are derived from GDP-mannose, a molecule that cannot cross the ER membrane (in part because it has no transporter in the ER membrane)³. The cell elegantly circumvents this obstacle by synthesizing dolichylmonophosphate-linked mannose on the outer side of the ER membrane and then somehow flipping this molecule, too, into the ER. Glucose is transported into the ER by a similar mechanism⁴ (Fig. 1).

The repeated 'somehows' in the last paragraph indicate that, although it has been known for over 15 years that lipid-linked sugars are flipped across the ER membrane, and despite repeated experiments, there have until now been no clues as to how the process takes place. Enter Helenius *et al.*¹. This group previously identified a mutant strain of budding yeast defective in N-glycosylation⁵. They found that this strain has a mutation in the *RFT1* gene, and their preliminary (still unpublished) results showed that *RFT1* encodes an essential protein — with several transmembrane regions — that localizes to the ER. The authors now have evidence of exactly what the Rft1 protein does.

Using some cunning genetic tricks, Helenius *et al.*¹ made a yeast strain in which expression of Rft1 could be repressed by removing glucose from the culture medium. In the absence of glucose, dolichylpyrophosphate-linked Man₃GlcNAc₂ accumulated at the expense of more complex lipid-linked oligosaccharides, and target proteins were underglycosylated. Through several elegant experiments, Helenius *et al.* discard the possibilities that this results from a lack of the enzyme that adds the sixth mannose unit, or from a decreased availability of dolichylmonophosphate-linked mannose.

The authors also assessed budding-yeast cells that lack Alg11, the enzyme responsible for adding the fifth mannose unit. These cells grow slowly and make severely underglycosylated glycoproteins. They also accumulate dolichylpyrophosphate-linked Man₃GlcNAc₂, which is moved inefficiently

Daedalus

Electric air drier

All modern plastic films have long-chain molecules. Stretching aligns them in the plane of the film, making the film stronger and highly impermeable to small molecules. Daedalus now wants a film with molecules oriented perpendicular to its plane. Langmuir–Blodgett films are like this, but they are normally too thin for the purposes Daedalus has in mind.

Langmuir–Blodgett films are made from a substance whose long-chain molecules float on a solvent such as water; their chains are vertical and nearly touch. When a substrate is passed through the water's surface, the film is transferred to the substrate. Many such passes build up a Langmuir–Blodgett film of numerous perpendicular layers. To make the films thicker, DREADCO chemists will lay them on a flexible substrate, perhaps paper or non-woven porous film. Many passes through a polymer solution will then be needed to build a thick Langmuir–Blodgett film. The frail product will be retained on the substrate, although it may be stripped off if it is strong enough. Its molecules will run across the film.

Unlike normal films, DREADCO's 'Langblofilm' will have little strength and will be highly permeable. With its molecular chains side-by-side across its whole width, it is in effect riddled with long, narrow, intermolecular holes lying between the chains. Small molecules will percolate through them. Daedalus recalls that in electro-osmosis a liquid is impelled along a narrow capillary by a voltage across its ends. This works best for capillaries of small radius, and for fluids of high permittivity. Water has a very high permittivity, so Langblofilm is an ideal pump for water.

One obvious use is air conditioning. Daedalus will evaporate a thin layer of aluminium on the film faces, and will lay it as a wallpaper on a porous wall. An electro-osmotic voltage applied across the faces will then suck water out of the air and into the film. At the other face it will be desorbed, and will diffuse through the porous wall to the outside. In this way the whole area of wall will dehumidify the room. Traditional air-conditioning systems, with their complex machines and pipework, will be totally outclassed. To suit popular taste, the DREADCO team may have to print floral patterns on their product, overlaying the aluminium electrode layer. This itself must be thin enough not to block the holes of the film, but weak electrostatics over a wide area will then take over from air-conditioning motors pumping on a tiny cross-section.

David Jones

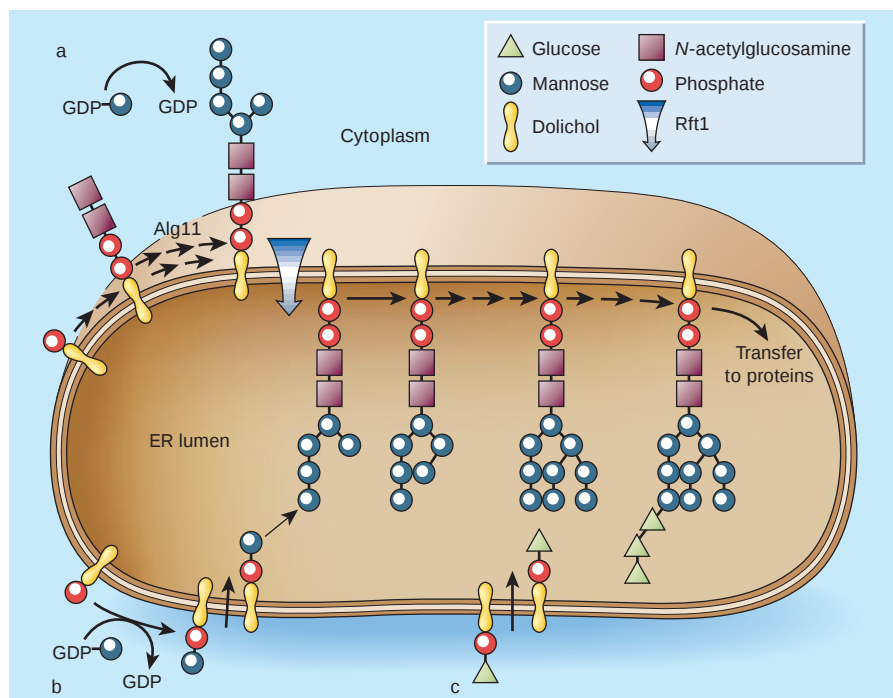


Figure 1 Cellular events that lead to the modification of proteins with sugar groups. **a**, The process starts with the sequential addition of seven monosaccharides (*N*-acetylglucosamine-phosphate, *N*-acetylglucosamine and five mannose units) to a lipid derivative (dolicholmonophosphate). This lipid-linked oligosaccharide is then flipped into the endoplasmic reticulum (ER), probably by the protein Rft1, as Helenius *et al.*¹ now show. Inside the ER, this oligosaccharide is further modified with four more mannose units and three glucose units before it can be attached to target proteins. **b**, **c**, The sources of mannose and glucose are, respectively, dolicholmonophosphate-linked mannose and dolicholmonophosphate-linked glucose, which are also flipped into the ER, by unknown mechanisms.

to the ER lumen where four more mannose units are added. Helenius *et al.*¹ overexpressed Rft1 in yeast lacking Alg11, and found that fewer glycoproteins were underglycosylated. Moreover, the cellular levels of lipid-linked $\text{Man}_3\text{GlcNAc}_2$ were reduced, but levels of lipid-linked $\text{Man}_7\text{GlcNAc}_2$ were increased. This oligosaccharide was later transferred to proteins. The implication of these and further results is that Rft1 is involved in moving polyprenol-bound sugars into the ER lumen. In these particular mutant cells, Rft1 moves $\text{Man}_3\text{GlcNAc}_2$ into the lumen. (In normal yeast, it would flip $\text{Man}_5\text{GlcNAc}_2$ inwards.) Significantly, *RFT1* has counterparts in all eukaryotic genomes sequenced to date, with the notable exception of *Plasmodium falciparum* — the malaria-causing parasite — which lacks *N*-glycosylation⁶.

Despite this advance there are still many longstanding unanswered questions. For example, is there a permanent physical association between the flipping protein (or proteins) and the polyprenol group? Such an association would presumably facilitate flipping, as dolichols are extremely large lipids (with up to 105 carbon atoms in mammalian cells) that are not expected to travel through the ER membrane as quickly as do phospholipids, the main cell-membrane components. In addition, what are the energy requirements involved in flipping a hydrophilic oligosaccharide into the ER?

What structural rearrangements occur in the flipping proteins? And do the flipping proteins interact with those that add sugar units to the oligosaccharide?

Finally, why can the membrane protein glucosidase I remove the outer glucose unit from the target-protein-linked oligosaccharide when the protein is still being synthesized, but not when the oligosaccharide is linked to the polyprenol? Does the flipping protein somehow protect this glucose? This is an important point, as $\text{Glc}_2\text{Man}_9\text{GlcNAc}_2$ is transferred to protein much less efficiently than the complete oligosaccharide. A defenceless lipid-linked oligosaccharide would undoubtedly lead to glycoprotein underglycosylation and misfolding.

Helenius *et al.*'s paper¹ does not answer any of these questions. But, rather like Ariadne in Greek mythology, it provides the beginning of a thread that will guide researchers out of the labyrinth of the flipping process.

Armando J. Parodi is at the Institute for Biotechnological Research, University of San Martin, CC 30, 1650 San Martin, Argentina. e-mail: aparodi@iib.unsam.edu.ar

1. Helenius, J. *et al.* *Nature* **415**, 447–450 (2002).
2. Parodi, A. J. *Annu. Rev. Biochem.* **69**, 69–95 (2000).
3. Abeijon, C. & Hirschberg, C. *Trends Biochem. Sci.* **17**, 32–36 (1992).
4. Kornfeld, R. & Kornfeld, S. *Annu. Rev. Biochem.* **54**, 631–664 (1985).
5. Ng, D. T., Spear, E. D. & Walter, P. *J. Cell Biol.* **150**, 77–88 (2000).
6. Davidson, E. A. & Gowda, D. C. *Biochimie* **83**, 601–604 (2001).

Obituary

Anil Agarwal (1947–2002)



Journalist and environmental campaigner

There can't be many science journalists who, having described the government of their country as a "persistent pest" in need of a "truly deadly pesticide", have been praised at their death by both the president and prime minister of that country as a national hero.

But Anil Agarwal, whose death on 2 January at the age of 54 ended a long fight against cancer, was no ordinary science journalist. A passionate and articulate campaigner for environmental justice, embracing issues ranging from water quality in rural India to the need for global controls on carbon emissions to limit climate change, Agarwal won respect from friends and enemies alike.

As the director of the Centre for Science and Environment in Delhi, a non-governmental organization that he created in 1980 on his return to India after working in Europe, Agarwal spearheaded the creation of a new form of activism. It was based on a commitment to sound scientific knowledge, combined with the idea that worries about the natural environment, and its destruction by the by-products of modern technologies, is not just a luxury of the industrialized world, but a major concern to the poor of developing countries too.

This idea is commonplace today, and forms the cornerstone of the World Summit on Sustainable Development, which takes place in Johannesburg later this year. But it was far from accepted 20 years ago, when concern for the environment was seen as incompatible with economic development. Similarly, as the founding editor of *Down to Earth*, a popular-science magazine that he created

in 1992 to promote the critical analysis of such issues, Agarwal pioneered a new form of science journalism. It is based on a commitment to expose the failures of government policy to protect the health of individuals, and to campaign, for example, for cleaner technologies, such as the use of compressed natural gas for public transport in Delhi.

An intense man, who reminded many of one of his heroes, Mahatma Gandhi, Agarwal trained as an engineer at the Indian Institute of Technology in Kanpur, graduating in 1970. After a short spell in the commercial world, an interest in environmental issues led him to Europe for the United Nations Environment Conference in Stockholm in 1972. Articles written about the meeting and related issues, including a series of interviews with prominent European scientists, led to a number of journalistic appointments in India, and subsequently to a feature-writing position with the International Institute for Environment and Development in London. His journalistic career developed steadily, and he regularly contributed to publications such as *Nature* and *New Scientist*.

But his interest in the relationship between poverty and environmental protection also intensified. Given Agarwal's commitment, it was little surprise when he returned to India to try to make his mark on this issue in his home country. There followed more than 20 years of active journalism and vigorous campaigning that led to changes in many areas of government policy. These ranged from legislation on air pollution to India's position in international

negotiations on the implementation of the Kyoto Protocol limiting carbon emissions into the atmosphere.

Much of Agarwal's work involved supporting grass-roots efforts to protect local communities against unscrupulous developers. For example, he championed the cause of the Chipko 'tree huggers' — women in a small Himalayan village who put their arms around trees to try to stop foresters cutting them down. Other efforts were more globally oriented. Agarwal was among the first to argue that concepts of social equity need to be integrated into international policies aimed at mitigating the harmful effects of human-induced climate change. Initially his was a relatively lone voice, but his arguments have had increasing influence on government policies in many developing countries. And his campaigning was not solely critical. For example, he argued vigorously for the widespread adoption of strategies for 'water-harvesting', a technique developed centuries earlier in parts of India, but neglected in many modern water-engineering projects.

Agarwal was awarded many prizes for his work combining care for the natural environment and social development. In 1979, for example, he was the first winner of the A. H. Boerma Award given by the Food and Agriculture Organization in Rome. And in 1987 he was elected to the Global 500 Honour Roll by the United Nations Environment Programme. He also pursued environmental causes at the international, as well as the national, level. Between 1984 and 1987 he was chair of the board of directors of the Environment Liaison Centre in Nairobi, and from 1988 to 1990 he was on the council of the World Resources Institute in Washington DC.

In the early 1990s, Agarwal was diagnosed with a rare form of non-Hodgkin's lymphoma that affected his sight and his brain. He wrote openly about his illness, castigating the lack of reliable statistics on cancer incidence in India, and raising questions about the extent to which his disease might have been caused by exposure to man-made chemicals in the environment.

India's president, K. R. Narayanan, described Agarwal's life and deeds as "a shining example of dedication to making mankind's onward advancement consistent with ecological protection". Agarwal's colleagues at the Centre for Science and Environment and elsewhere are carrying on with this task. But as they admit, his absence will make it more difficult.

David Dickson

David Dickson is at SciDev.Net, 11 Rathbone Place, London W1T 1HR, UK.
e-mail: david.dickson@scidev.net

Fertility after intact ovary transplantation

Frozen banking of whole organs for transplantation is starting to look feasible.

Transplantation surgery, which is limited by the supply and short-term viability of fresh donor organs, would be revolutionized if these could survive freezing, but early claims of cryopreservation¹ were never realized. Here we describe the successful transplantation in rats of ovaries, fallopian tubes and the upper segment of the uterus *en bloc* after storage in liquid nitrogen.

We anaesthetized adult female Lewis rats using sodium pentobarbital and removed the right ovary and the reproductive tract, with the ovarian vessels dissected to create short cuffs of aorta and vena cava^{2,3}; these preparations were held in solution for less than an hour before transplantation. Syngeneic recipients were anaesthetized, heparinized and oophorectomized. The aorta and vena cava below the left renal artery were dissected for end-to-side aortic–aorta and veno–venous anastomoses using continuous suturing; the uterus was joined to the corresponding host segment. After paired operations lasting for 2 h, circulation was restored to donor organs (for further details, see supplementary information). All surviving recipients stayed healthy.

We used eight fresh organs for transplantation and also seven treated ones that we had perfused for 30 min at 0.35 ml min⁻¹ with M2 medium containing 0.1 M fructose and increasing concentrations of dimethylsulphoxide (0–1.5 M). Following a protocol previously used for sliced ovarian tissue⁴, we cooled the treated organs slowly in 5-ml cryovials (Nalgene), inducing ice nucleation at –7 °C in the modified chamber of an automated freezer (CryoLogic). After overnight storage in liquid nitrogen, we rapidly thawed the vials and removed the cryoprotectant by perfusion with a reversed concentration gradient (for further details, see supplementary information).

We monitored ovarian function by taking vaginal smears (see supplementary information) and by pairing some animals with stud males four weeks later. Ten weeks after surgery, we collected blood immediately after euthanasia to measure serum concentrations of follicle-stimulating hormone (FSH) and oestradiol-17 β by radioimmunoassay and chemiluminescence, respectively. We prepared organs for histology and to assess ovarian follicle reserves.

All animals with fresh transplants reinitiated oestrus. They had normal numbers of ovarian follicles, including graafian stages (Fig. 1a), and uterine weights and FSH levels were similar to those of controls (9.7 ± 0.5 and 9.1 ± 0.4 ng ml⁻¹, respectively). Four animals with cryopreserved grafts (57%) had

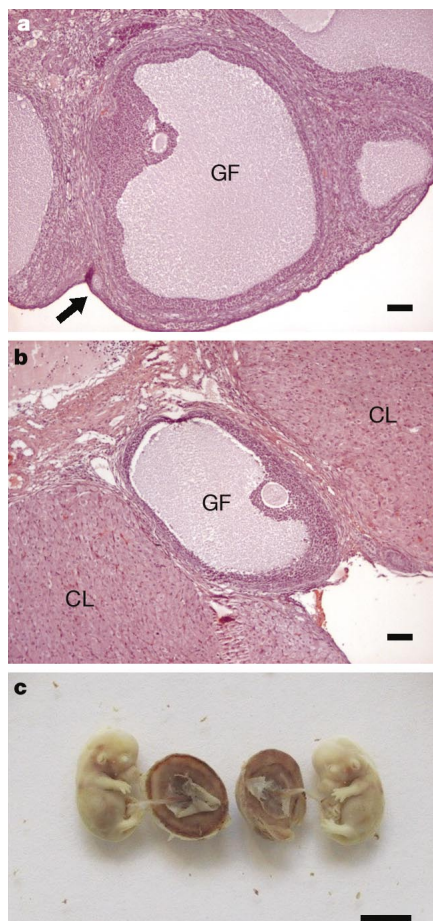


Figure 1 Transplantation success of ovaries and reproductive tract in rats. **a**, Fresh-ovary transplants had growing follicles (arrow) and mature graafian follicles (GF). **b**, Cryopreserved-ovary transplants were also follicular and contained corpora lutea (CL). **c**, Two fetuses conceived by one animal that had received a frozen–thawed transplant. Scale bars: 0.05 mm (**a**, **b**) and 5 mm (**c**).

follicular ovaries, and their corpora lutea indicated recent ovulation (Fig. 1b). One animal was pregnant, with two healthy fetuses implanted on either side of the uterine anastomosis (Fig. 1c). This group had higher

serum FSH levels (21.3 ± 6.2 ng ml⁻¹), fewer follicles and lower oestradiol levels and uterine weights, but none was in the castrate range. Tubal and uterine morphology were indistinguishable from unoperated controls.

These results are encouraging, but indicate that ovarian function is compromised to some extent by freezing, perhaps because of intravascular ice formation — a problem that has previously frustrated attempts to cryopreserve kidneys⁵. Advances in vitrification may overcome this problem if the chemical toxicity of additives can be minimized⁶. Frozen banking of reproductive organs could eventually be useful in breeding from endangered species and as a fertility option for women and children who have undergone sterilizing chemotherapy^{7,8}. Our success with ovarian transplants should stimulate investigation into the improvement of cryopreservation techniques for other organs.

Xiang Wang*, **Huifang Chen***, **Hang Yin†**,
S. Samuel Kim‡, **Seang Lin Tan†**,
Roger G. Gosden†

*Laboratory of Experimental Surgery, Centre Hospitalier de l'Université de Montréal, Notre-Dame Hospital, Montreal, Quebec H2L 4M1, Canada

†Department of Obstetrics and Gynecology, McGill University, Royal Victoria Hospital, Montreal, Quebec H3A 1A1, Canada
e-mail: gosdenrg@evms.edu

‡Department of Obstetrics and Gynecology, University of Washington School of Medicine, Seattle, Washington 98105, USA

- Guttman, F. M. *et al. Cryobiology* **14**, 559–567 (1977).
 - Lee, S. J. *Res. Inst. Sci. Korea* **6**, 112 (1974).
 - Cornier, E., Sibella, P. & Chatelet, F. J. *Gynecol. Obstet. Biol. Reprod.* **14**, 567–573 (1985).
 - Gosden, R. G. *et al. Hum. Reprod.* **9**, 597–603 (1994).
 - Jacobsen, I. A. *et al. Cryobiology* **21**, 637–653 (1984).
 - Song, Y. C. *et al. Nature Biotechnol.* **18**, 296–299 (2000).
 - Pictou, H. M., Kim, S. S. & Gosden, R. G. *Br. Med. Bull.* **56**, 603–615 (2000).
 - Radford, J. A. *et al. Lancet* **357**, 1172–1175 (2001).
- Supplementary information** accompanies this communication on Nature's website (www.nature.com).

Competing financial interests: declared none.

Chemistry

Amplification by compartmentalization

Autocatalysis and chemical amplification are properties of living systems that can lead to increased responsiveness and to self-replication. Here we describe a synthetic system in which a unique form of reagent compartmentalization gives rise to nonlinear kinetics that are subject to the precise size- and shape-

selectivity of the host. The reactivity is reminiscent of autocatalytic behaviour¹, in which there is no direct contact between reagents and products, and our approach offers a general way to impose complex chemical behaviour onto synthetic systems.

Reversibly formed hosts, such as capsule **1** in Fig. 1 (ref. 2), temporarily surround smaller guest species and effectively divide a reaction milieu into two domains: reagents are either free in solution ('on'), where they react normally, or are encapsulated ('off'), where they are unreactive. The exchange of

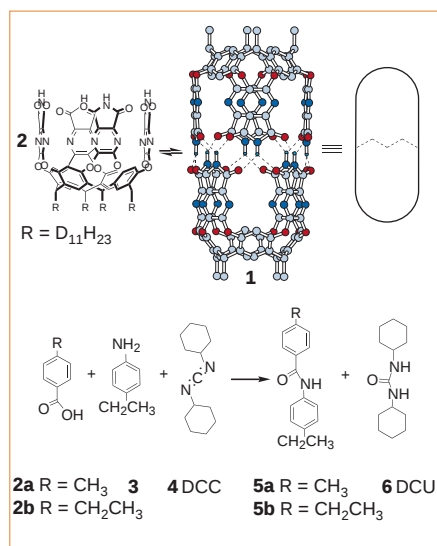


Figure 1 The dimeric capsule **1** encapsulates molecules of the proper size and shape and inhibits their reactivity. Reagent **4** and products **5a** and **6** are good guests, but the slightly larger product anilide **5b** is not. DCC, dicyclohexylcarbodiimide; DCU, dicyclohexylurea.

guest species between these environments is thus a means of controlling reactivity.

Consider the reaction of *p*-toluic acid (molecule **2a** in Fig. 1) or *p*-ethylbenzoic acid, (**2b**) with *p*-ethylaniline (**3**) and dicyclohexylcarbodiimide (DCC; **4**). When 5 mmol each of **2a** or **2b**, 10 mmol of **3** and 1 mmol of **4** are mixed at 295 K in mesitylene, the almost identical anilides **5a** and **5b** are formed at almost identical rates, together with dicyclohexylurea (**6**) and *N*-acylurea byproducts.

These initial rates are much lower in the presence of stoichiometric amounts of **1**, which encapsulates **4** and ablates its reactivity. The initial, equilibrium concentration of free **4** is too low to be seen by nuclear magnetic resonance spectroscopy, and the reaction of **2a** or **2b** with the aniline proceeds only through these trace amounts of **4** in solution. In contrast to the kinetics in the absence of **1**, the reaction of the shorter acid **2a** is now much faster than that of the longer **2b** (Fig. 2). Moreover, addition of the shorter product **5a** accelerates the initial rate of the reaction involving **2a**, but addition of the only slightly longer product anilide **5b** has no effect.

This behaviour results from feedback loops in a self-regulating reaction cycle (Fig. 2a). The urea molecule **6** and toluic anilide (**5a**) — the products of the reaction — are both good guests for the host **1** (relative binding affinities: $5a \sim 6 > 4$). Once formed, the products displace DCC from the capsule into the bulk solvent, where it can react with the acid. A single molecule of DCC reacts to yield one molecule each of urea and anilide, which each displace further DCC molecules, leading to the production of more urea and anilide in a chain

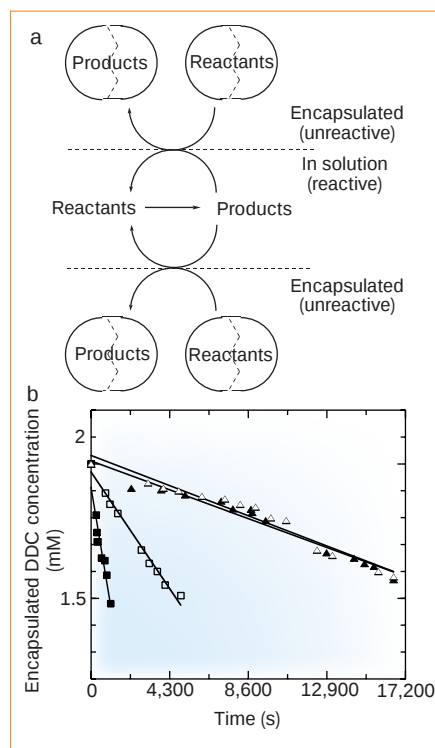


Figure 2 Reaction of compartmentalized reagents. **a**, For the reaction shown in Fig. 1, products **5a** and **6** displace reagent **4** from the capsule interior, creating a feed-forward reaction cycle. When acid **2b** is used, the slightly larger product anilide **5b** is a poor guest, and half of the reaction cycle is not operative. **b**, Dicyclohexylcarbodiimide (DCC) released from the capsule interior as a function of time for the reactions shown in Fig. 1. Initial concentrations: **4**, 2 mM; **1**, 2 mM; **3**, 10 mM; **2a**, 2 mM (open squares) or 2 mM with 0.25 mM equivalents of amide **5a** (filled squares); **2b**, 2 mM (open triangles) or 2 mM with 0.25 mM equivalents of amide **5b** (filled triangles). Errors in integrations and concentrations are within 5%; straight lines are shown for guidance.

reaction. The capsule does not influence the reaction between acid and DCC — it merely limits the rate at which reagents encounter each other, giving rise to kinetics that accelerate with the increased concentration of free 'reactive' DCC.

The reaction of **4** with the very similar **2b** occurs with different kinetics because the product molecule *p,p'*-diethylbenzanilide (**5b**) cannot fit into the capsule³. Only one good guest — the urea molecule **6** — is generated in this instance, and no further reagents are released; the kinetics are not accelerated.

This behaviour, although easily understood, is difficult to classify. The kinetics have a sigmoidal character that depends on product formation, but the reaction is not autocatalytic as it has no true catalyst. Moreover, unlike templated autocatalysis, no direct contact is required between reagents and products — an advantage in that catalyst inhibition due to self-complementary products is not an obstacle. The nonlinear kinetics are viewed more correctly as an emergent property of the system as a whole, rather than of specific

molecules within the system.

The combination of compartmentalization and molecular recognition does, however, result in chemical amplification, which is useful for increased sensitivity and self-replication. Other compartmentalized chemistry has similar potential, and product formation influences subsequent reactivity in zeolites⁴. The inherent benefits of compartmentalization are thought to be an important, if not essential, characteristic of living systems (for example, see ref. 5). As long as the relative timescales of guest exchange and reaction are appropriate, our approach can be generalized to a range of reactions and even to complex but well-defined chemical systems.

Jian Chen*, **Steffi Körner***,
Stephen L. Craig†, **Dmitry M. Rudkevich‡**,
Julius Rebek Jr*

*Skaggs Institute for Chemical Biology, Scripps Research Institute, 10550 North Torrey Pines Road, La Jolla, California 92037, USA

e-mail: jrebek@scripps.edu

†Department of Chemistry, Duke University, Durham, North Carolina 27708-0346, USA

‡Department of Chemistry and Biochemistry, University of Texas at Arlington, Arlington, Texas 76019-0065, USA

- Robertson, A., Sinclair, A. J. & Philp, D. *Chem. Soc. Rev.* **29**, 141–152 (2000).
- Heinz, T., Rudkevich, D. M. & Rebek, J. *Jr Nature* **394**, 764–766 (1998).
- Körner, S. K., Tucci, F. C., Rudkevich, D. M., Heinz, T. & Rebek, J. *Jr Chem. — A Eur. J.* **6**, 187–195 (1999).
- Turro, N. J., Han, N., Lei, X., Fehner, J. R. & Abrams, L. *J. Am. Chem. Soc.* **117**, 4881–4893 (1995).
- Francastel, C., Schubeler, D., Martin, D. I. K. & Groudine, M. *Nature Rev. Mol. Cell Biol.* **1**, 137–143 (2000).

COMMUNICATIONS ARISING

Palaeoclimatology

CO₂ and the end-Triassic mass extinction

The end of the Triassic period was marked by one of the largest and most enigmatic mass-extinction events in Earth's history¹ and, with few reliable marine geochemical records, terrestrial sediments offer an important means of deciphering environmental changes at this time. Tanner *et al.*² describe an isotopic study of Mesozoic fossil soils which suggests that the atmospheric concentration of carbon dioxide (*p*CO₂) across the Triassic–Jurassic boundary was relatively constant (within 250 p.p.m.v.), but this is inconsistent with high-resolution evidence from the stomatal characters of fossil leaves³. Here I show that the temporal resolution of the fossil-soil samples may have been inadequate for detecting a transient rise in *p*CO₂. I also show that the fossil-soil data are consistent with a large increase in *p*CO₂ across the Triassic–Jurassic boundary

when variations in the stable carbon isotope (denoted as $\delta^{13}\text{C}$) in terrestrial plant leaves are taken into account. These factors suggest that the linkage between $p\text{CO}_2$, global warming and the end-Triassic mass extinction remains intact.

The stability of the atmospheric carbon reservoir across the Triassic–Jurassic boundary is inferred from $p\text{CO}_2$ estimates derived using measurements of the $\delta^{13}\text{C}$ composition of pedogenic carbonates, ranging over 20 million years (between the Late Triassic and Early Jurassic), and a palaeosol $p\text{CO}_2$ barometer⁴. A temporal resolution of 20 million years is, however, lower than in the study of fossil leaves³ and so is unlikely to be compatible with detecting the same transient carbon-cycle event. Moreover, a well-characterized feature of the palaeosol $p\text{CO}_2$ barometer is its sensitivity, not only to the concentration of CO_2 in the soil, as acknowledged by Tanner *et al.*², but also to the isotope composition of terrestrial organic matter ($\delta^{13}\text{C}_{\text{OM}}$)^{4,5}. It is therefore necessary to measure $\delta^{13}\text{C}_{\text{OM}}$ coexisting with soil carbonates^{4,5}, but unfortunately, under the very conditions in which palaeosol carbonates form (well-drained soils from arid and semi-arid regions), soil organic matter is rarely preserved, so this feature of the method is often poorly constrained.

Tanner *et al.* calculate $p\text{CO}_2$ by applying a single $\delta^{13}\text{C}_{\text{OM}}$ value because they lacked organic materials with a stratigraphic resolution that corresponds to all of their soil samples. For the interval across the Triassic–Jurassic boundary, however, at least one land-plant stable-carbon-isotope record is available³, providing a means to account for natural variations in $\delta^{13}\text{C}_{\text{OM}}$ at this time. This negative isotopic excursion has subsequently been recorded in organic matter from the Queen Charlotte Islands in British Columbia⁶ and from St Audries Bay, UK⁷, as

well as in marine carbonates and organic matter from Hungary⁸, indicating that it could be a general geochemical fingerprint of the carbon-cycle perturbation across the Triassic–Jurassic boundary.

Calculation of palaeo- $p\text{CO}_2$ levels using the mean isotopic ratios of Triassic and Jurassic palaeosol carbonates determined by Tanner *et al.*, but constrained using stratigraphically detailed terrestrial-plant $\delta^{13}\text{C}_{\text{OM}}$ records from Jameson Land, east Greenland³, indicates a substantial increase across the Triassic–Jurassic boundary (Fig. 1a). Within the margins of uncertainty of the palaeosol approach, this increase in $p\text{CO}_2$ is consistent with the original $p\text{CO}_2$ reconstruction based on fossil stomata of *Ginkgo* cuticles³ (Fig. 1b) and equates to a rise of around 1,000 p.p.m.v. The rise is more strongly expressed if CO_2 -related global warming is allowed to influence the fractionation of soil CO_2 in the reaction-diffusion model in a manner that is consistent with theory (Fig. 1a). Both the stomatal and palaeosol reconstructions (Fig. 1) show a marked $p\text{CO}_2$ increase that is broadly coincident with the emplacement of the Central Atlantic Magmatic Province (CAMP) flood basalts⁹.

Neither the palaeosol nor the stomatal approach to estimating $p\text{CO}_2$ variations in deep time are certain, but both provide self-consistent results for the Palaeozoic⁴ era and the mass-extinction event across the Triassic–Jurassic boundary (Fig. 1). Such self-consistency argues against strongly coordinated ocean–atmosphere buffering of CO_2 from volcanic outgassing during the end of the Triassic, possibly because of the geologically short duration (about 0.5 Myr)¹⁰ and high intensity of CAMP activity. Precise dating of the proposed CO_2 excursion is critical in this context.

Moreover, it is hard to envisage a mechanism other than extreme, CO_2 -forced

global warming that can explain the widespread selective floral extinctions and turnovers that are evident in the fossil record at this time. The proposed transgressive–regressive change in sea level² would be unlikely to provoke such a marked response from the terrestrial biota. The mechanisms that drive mass extinction in the marine and terrestrial realms warrant urgent investigation in the face of the current exponential rise in the Earth's atmospheric CO_2 concentration.

David Beerling

Department of Animal and Plant Sciences,
University of Sheffield, Sheffield S10 2TN, UK
e-mail: d.j.beerling@sheffield.ac.uk

1. Hallam, A. & Wignall, P. B. *Mass Extinctions and their Aftermath* (Oxford Univ. Press, Oxford, 1997).
2. Tanner, L. H., Hubert, J. F., Coffey, B. P. & McInerney, D. P. *Nature* **411**, 675–677 (2001).
3. McElwain, J. C., Beerling, D. J. & Woodward, F. I. *Science* **285**, 1386–1390 (1999).
4. Royer, D. L., Berner, R. A. & Beerling, D. J. *Earth-Sci. Rev.* **54**, 349–392 (2001).
5. Cerling, T. E. *Glob. Biogeochem. Cycles* **6**, 307–314 (1992).
6. Ward, P. D. *et al. Science* **292**, 1148–1151 (2001).
7. Hesselbo, S. P. *et al. Geology* (in the press).
8. Pálffy, J. *et al. Geology* **29**, 1047–1050 (2001).
9. Marzoli, A. *et al. Science* **284**, 616–618 (1999).
10. Hames, W. E., Renne, P. R. & Ruppel, C. *Geology* **28**, 859–862 (2000).

Triassic–Jurassic atmospheric CO_2 spike

I question the claim by Tanner *et al.*¹ that atmospheric CO_2 levels remained constant across the Triassic–Jurassic boundary on the grounds of problems with stratigraphic completeness and contamination with atmospheric methane. Because methanogenic CH_4 has a light isotope composition and oxidizes readily to CO_2 , methane–clathrate dissociation and oxidation events cannot be detected by palaeobarometers that use the carbon-isotope composition of palaeosol carbonate.

The palaeosol isotopic data of Tanner *et al.* are not presented in the context of measured sections, nor is there any other indication of stratigraphic completeness that is suitable for identification of a transient greenhouse effect. It is therefore unsurprising that the few sampled palaeosols missed the brief CO_2 greenhouse effect inferred from the stomatal index of fossil leaves in three beds within two continuous sequences in Sweden and Greenland². The duration of the isotopic excursion that occurred at the same time as this transient CO_2 greenhouse effect was probably no more than 500,000 years³. This slim target is easy to miss when examined from a distance of 200 million years.

A more serious problem with isotopic palaeobarometers of CO_2 is the compromising effect of massive dissociation events

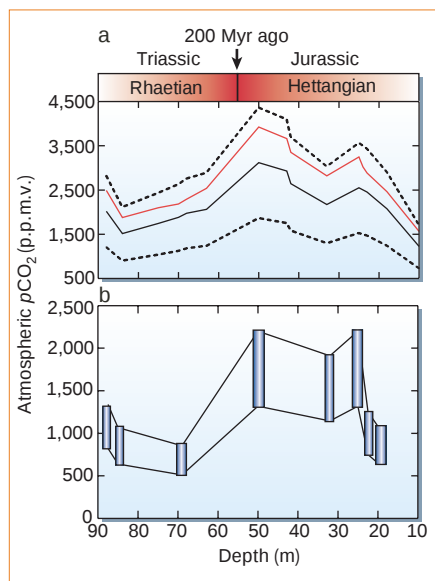


Figure 1 Palaeo-atmospheric $p\text{CO}_2$ changes across the Triassic–Jurassic boundary 208 million years ago. **a**, Atmospheric $p\text{CO}_2$ changes were calculated using the carbon-isotope composition of North American fossil soils² and land-plant organic-matter records from Jameson Land, east Greenland³, using a standard diffusion model⁵, where $S(z)$ (the concentration of CO_2 contributed by biological respiration) is 5,000 p.p.m.v. The horizontal axis represents the depth of successive fossil-plant-bearing beds of the Cape Stewart formation from Jameson Land. $p\text{CO}_2$ values were calculated using a constant Triassic palaeosol carbonate value up to the Triassic–Jurassic boundary and thereafter a constant Jurassic palaeosol value (both from ref. 2). Red line indicates the effects of CO_2 -driven global warming on the temperature dependence of soil CO_2 fractionation^{4,5}, with $S(z) = 5,000$ p.p.m.v. only. Upper and lower broken lines denote the variability in these estimates by varying $S(z)$ between 3,000 and 7,000 p.p.m.v. **b**, Atmospheric $p\text{CO}_2$ changes reconstructed from the stomatal characteristics of fossil leaves from Jameson Land³. Vertical bars denote the upper and lower range for any given depth calculated using this technique.

from methane hydrate reservoirs, because CH_4 oxidizes in the atmosphere within 7–24 years to CO_2 , which retains a very depleted carbon-isotope composition (typically -60% , but as low as -110% $\delta^{13}\text{C}_{\text{org}}$)⁴. For example, unusually depleted carbon-isotope values for Early Triassic organic matter can be taken as indications of a methane-dissociation event⁵. An Early Triassic CO_2 greenhouse effect is shown by the very low stomatal index of fossil seed ferns⁶, but pedogenic carbonate isotopic palaeobarometers indicate low atmospheric CO_2 levels in the Early Triassic⁷. The pedogenic–isotopic palaeobarometer was not designed for methanogenic isotopic compositions, which compromise this palaeobarometer during several greenhouse transients⁶, and perhaps also at the Triassic–Jurassic boundary.

Gregory J. Retallack

Department of Geological Sciences, University of Oregon, Eugene, Oregon 97403-1272, USA
e-mail: gregr@darkwing.uoregon.edu

1. Tanner, L. H., Hubert, J. F., Coffey, B. P. & McInerney, D. P. *Nature* **411**, 675–677 (2001).
2. McElwain, J. C., Beerling, D. J. & Woodward, F. I. *Science* **285**, 1386–1390 (1999).
3. Ward, P. D. *et al. Science* **292**, 1148–1151 (2001).
4. Whiticar, M. J. in *Atmospheric Methane* (ed. Khalil, M. A. R.) 63–85 (Springer, Berlin, 2000).
5. Krull, E. S. & Retallack, G. J. *Bull. Geol. Soc. Am.* **112**, 1459–1472 (2000).
6. Retallack, G. J. *Nature* **411**, 287–290 (2001).
7. Ekart, D. D., Cerling, T. E., Montañez, I. P. & Tabor, N. J. *Am. J. Sci.* **299**, 805–827 (1999).

Tanner replies — Both Beerling and Retallack question our conclusion that atmospheric CO_2 levels remained relatively stable across the Triassic–Jurassic boundary, partly on the basis of insufficient stratigraphic resolution. However, the stratigraphy of the formations we studied is well known¹. The Lower Jurassic McCoy Brook Formation of the Fundy basin overlies the North Mountain Basalt, which was extruded during the initial stages of volcanism in the Central Atlantic Magmatic Province (CAMP), and so this formation post-dates the main eruptive episode. Palaeosols in this formation occur within 10 m of the formation base; the age of these palaeosols is therefore constrained by the basalt to within several hundred thousand years of the Triassic–Jurassic boundary, which is located several metres below the basalt. Because the duration of the eruptions of CAMP volcanics, which occurred in several pulses, has been established as roughly 600,000 years², palaeosol formation in the McCoy Brook Formation is contemporaneous with the latter stages of the CAMP eruptions. These palaeosols may therefore be expected to record the cumulative effects of the eruptions.

Our palaeo- CO_2 values are calculated from the diffusion-reaction model, which

requires measurements or assumptions for a variety of factors that control soil CO_2 , not least the isotopic composition of plant-derived organic matter ($\delta^{13}\text{C}_{\text{OM}}$)^{3,4}. Ideally, organic matter in the palaeosol that contains the carbonate is used to obtain this value because the $\delta^{13}\text{C}$ of C_3 plants is known to vary significantly in contemporaneous soils from differing climatic regimes³. Unfortunately, the McCoy Brook Formation, from which the Lower Jurassic carbonate samples were obtained, lacks well-preserved plant material.

The only record of $\delta^{13}\text{C}_{\text{OM}}$ across the Triassic–Jurassic boundary available at the time of our calculations provides data for locations in eastern Greenland and southern Sweden⁵, but the negative $\delta^{13}\text{C}$ excursion at the boundary shown in the eastern Greenland data, represented primarily by a single data point, is not apparent in the southern Sweden data set. Moreover, the sediments that contain plant fossils at the eastern Greenland location accumulated under the influence of a significantly more humid climate than existed in the Fundy basin¹, a fact that renders isotope data from this location inapplicable to the interpretation of data derived from the semi-arid palaeosols.

For these reasons, we chose not to use these data, and instead assumed a single value for both the Late Triassic and Early Jurassic that is consistent with published measures of organic matter in Upper Triassic formations³ that were deposited under climatic conditions similar to those of the studied formations. This potential source of inaccuracy may ultimately be resolved only by location and analysis of organic material in the McCoy Brook Formation.

The fossil stomatal evidence of Beerling and Retallack seems to indicate a several-fold increase in atmospheric $p\text{CO}_2$ across the Triassic–Jurassic boundary^{5,6}. The use of stomatal indices for calculation of palaeo- CO_2 levels is based on experiments in which modern plants were grown at $p\text{CO}_2$ values of up to twice present levels⁷, but calls for an extrapolation of the experimental $p\text{CO}_2$ values to the much higher palaeo- CO_2 values interpreted for the past. The use of these indices also requires the assumption that the floral response to these conditions was quantitatively the same 200 million years ago as it would be today. This approach may therefore also generate inaccuracies.

As mentioned by Beerling, data from the marine realm that indicate a significant negative carbon-isotope excursion at the Triassic–Jurassic boundary shed new light on the extinction problem⁸, demonstrating an intense but short-lived perturbation of the global carbon cycle of a magnitude that is not consistent with volcanic outgassing. This is supported by a simple mass-balance

calculation, on the basis of the largest estimate of the volume and volatile content of intrusions of the CAMP⁹, showing that outgassing during the eruptions produced 5.6×10^{16} mol CO_2 in total, an amount that is equivalent to that in the modern atmosphere.

This volume is insufficient to drive the isotopic excursion of marine carbonates⁸ and is unlikely to have affected the atmosphere of the early Mesozoic era, given the high $p\text{CO}_2$ of the Late Triassic¹. Rapid release of seafloor methane hydrate, as suggested by Retallack, has been implicated in other extinction events, and this release may have been triggered by the effects of CAMP volcanism. Such a release, which is consistent with the magnitude of the isotopic excursion, need not have resulted in greatly increased atmospheric $p\text{CO}_2$ as suggested by Retallack, as evidence exists that much of the methane released by this process would be oxidized within the water column, resulting in a brief interval of ocean anoxia¹⁰ and widespread extinction.

Our conclusion stands: the isotope compositions of pedogenic carbonates fail to indicate a substantial increase in atmospheric $p\text{CO}_2$ as a result of the CAMP eruptions. We are all in agreement that although existing methods of estimating palaeo- CO_2 are inexact, their validity is not mutually exclusive, and also that the cause of the end-Triassic extinction event is uncertain, with environmental degradation resulting from the CAMP volcanism probably being involved. Acquisition of new data by both methods from other locations should resolve this uncertainty; better time resolution will constrain the relationships between the abrupt marine extinctions¹¹, the possibly asynchronous floral turnover¹² and the duration of the CAMP eruptions, which may have lasted for several million years¹³.

Lawrence H. Tanner

Department of Geography and Geosciences, Bloomsburg University, Bloomsburg, Pennsylvania 17815, USA
e-mail: ltann@bloomu.edu

1. Tanner, L. H., Hubert, J. F., Coffey, B. P. & McInerney, D. P. *Nature* **411**, 675–677 (2001).
2. Olsen, P. E., Schlische, R. W. & Fedosh, M. S. *The Continental Jurassic* (ed. Morales, M.) 11–22 (Mus. North. Arizona, Flagstaff, Arizona, 1996).
3. Cerling, T. E. *Spec. Publ. Int. Ass. Sedimentol.* **27**, 43–60 (1999).
4. Ekart, D. D., Cerling, T. E., Montañez, I. P. & Tabor, N. J. *Am. J. Sci.* **299**, 805–827 (1999).
5. McElwain, J. C., Beerling, D. J. & Woodward, F. I. *Science* **285**, 1386–1390 (1999).
6. Retallack, G. J. *Nature* **411**, 287–290 (2001).
7. Beerling, D. *et al. J. Exp. Bot.* **49**, 1603–1607 (1998).
8. Pálffy, J. *et al. Geology* **29**, 1047–1050 (2001).
9. McHone, J. G. *Geol. Soc. Am. Abs. Prog.* **32**, 58–59 (2000).
10. Dickens, G. *Geochem. Geophys. Geosys.* **2**, 2000GC000131 (2001).
11. Ward, P. D. *et al. Science* **292**, 1148–1151 (2001).
12. Pálffy, J. *et al. Geology* **28**, 39–42 (2000).
13. Marzoli, A. *et al. Science* **284**, 616–618 (1999).

Antimicrobial peptides of multicellular organisms

Michael Zasloff*

Department of Biochemistry and Biophysics, University of Pennsylvania School of Medicine, Philadelphia, Pennsylvania 19104-6059, USA

Multicellular organisms live, by and large, harmoniously with microbes. The cornea of the eye of an animal is almost always free of signs of infection. The insect flourishes without lymphocytes or antibodies. A plant seed germinates successfully in the midst of soil microbes. How is this accomplished? Both animals and plants possess potent, broad-spectrum antimicrobial peptides, which they use to fend off a wide range of microbes, including bacteria, fungi, viruses and protozoa. What sorts of molecules are they? How are they employed by animals in their defence? As our need for new antibiotics becomes more pressing, could we design anti-infective drugs based on the design principles these molecules teach us?

Antimicrobial peptides are evolutionarily ancient weapons. Their widespread distribution throughout the animal and plant kingdoms suggests that antimicrobial peptides have served a fundamental role in the successful evolution of complex multicellular organisms. Despite their ancient lineage, antimicrobial peptides have remained effective defensive weapons, confounding the general belief that bacteria, fungi and viruses can and will develop resistance to any conceivable substance. Antimicrobial peptides target a previously under-appreciated 'microbial Achilles heel', a design feature of the microbial cellular membrane that distinguishes broad species of microbes from multicellular plants and animals. The insights provided by this large body of research have spawned considerable commercial effort to create new classes of anti-infective therapeutics.

A diversity of peptides

The diversity of antimicrobial peptides discovered is so great that it is difficult to categorize them except broadly on the basis of their secondary structure. (An online catalogue of all reported molecules, now about 500, can be found at <http://www.bbcm.univ.trieste.it/~tossi/antimic.html>). The fundamental structural principle underlying all classes is the ability of the molecule to adopt a shape in which clusters of hydrophobic and cationic amino acids are spatially organized in discrete sectors of the molecule ('amphipathic' design) (Fig. 1). Linear peptides, such as the silk moth's cecropin¹ and the African clawed frog's magainin², adopt this organization only when they enter a membrane, whereupon they assume an amphipathic α -helical secondary structure³. Frog species of the genus *Rana* modify this design by adding a single loop formed by a disulphide bond at the carboxy end⁴. Peptides such as bactenecin⁵ and defensins⁶ use a relatively rigid anti-parallel β -sheet constrained by disulphide bonds as the framework, around which segregated patches of cationic and hydrophobic residues are organized. A large family of linear peptides characterized by a predominance of one or two amino acids, such as the tryptophan-rich indolicidin of the cow neutrophil⁷ and the proline-arginine-rich PR39 (ref. 8) of the pig neutrophil, segregate hydrophobic and hydrophilic side chains around an extended peptide scaffold in the setting of the membrane (Table 1). Most multicellular organisms express a cocktail comprising multiple peptides from several of these structural classes within their 'defensive' tissues.

All antimicrobial peptides are derived from larger precursors,

including a signal sequence. Post-translational modifications include proteolytic processing, and in some cases glycosylation⁹, carboxy-terminal amidation and amino-acid isomerization (reviewed in ref. 4), and halogenation¹⁰. A rather complex modification involves the cyclization of two short peptides leading to the fully circular θ -defensin isolated from neutrophils of *Rhesus* (macaque) monkeys¹¹. Some peptides are derived by proteolysis from larger proteins, such as buforin II from histone 2A (ref. 12) and lactoferricin from lactoferrin¹³.

The diversity of sequences is such that the same peptide sequence is rarely recovered from two different species of animal, even those closely related, be they insects, frogs or mammals (exceptions include peptides cleaved from highly conserved proteins, such as buforin II). However, both within the antimicrobial peptides from a single species, and even between certain classes of different peptides from diverse species (reviewed in ref. 4), significant conservation of amino-acid sequences can be recognized in the preproregion of the precursor molecules; the design suggests that constraints exist on the sequences involved in the translation, secretion or intracellular

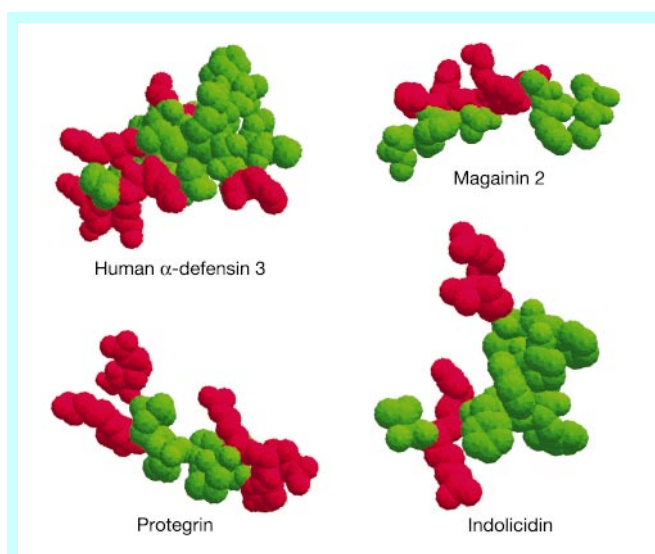


Figure 1 Clustering of cationic and hydrophobic amino acids into distinct domains in several antimicrobial peptides of different structural classes. This 'amphipathic' design is evident in many, but not all, antimicrobial peptides. Red, basic (positively charged) amino acids; green, hydrophobic ('oily') amino acids. Other amino acids are not shown. Magainin is depicted in its α -helical configuration.

* Present address: Georgetown University, Medical Center, 4000 Reservoir Road NW, Washington DC 2007-2197, USA

trafficking of this class of membrane-disruptive peptide. This feature is dramatically illustrated by the cathelicidins (reviewed in ref. 14).

Why does diversity arise? Because single mutations can dramatically alter the biological activity of each peptide, the diversity probably reflects the species' adaptation to the unique microbial environments that characterize the niche occupied, including the microbes associated with acceptable food sources (reviewed in refs 4 and 15). It seems reasonable to speculate that an individual could find itself in the midst of microbes against which the peptides of its species were ineffective; although the individual might suffer, the species itself could survive through emergence of individuals expressing beneficial mutations. Adaptive immunity, through its plasticity, permits a species to empower individuals to explore new environments and avail themselves of new food sources. However, compared with the equipment of the innate system, such as antimicrobial peptides, the effectors of adaptive immunity are more costly to maintain and slower to respond to assault¹⁵.

With respect to the diversity created in the synthetic laboratory, almost all active molecules are composed of hydrophilic, hydrophobic and cationic amino acids arranged in a molecule that can organize into an amphipathic structure (reviewed in ref. 16). Natural peptides composed of all D-amino acids, in place of L-amino acids, typically retain full antibiotic potency while exhibiting expected resistance to enzymatic proteolysis¹⁶. Short linear or cyclic amphiphilic peptides that contain both L- and D-amino acids can be generated with various degrees of selectivity and antimicrobial potency^{17,18}. Recently, protease-resistant antimicrobial peptides composed of β -amino acids have been constructed^{19,20}.

Mechanism

Antimicrobial peptides have targeted a surprising but clearly fundamental difference in the design of the membranes of microbes and multicellular animals, best understood for bacterial targets. Bacterial membranes are organized in such a way that the outermost

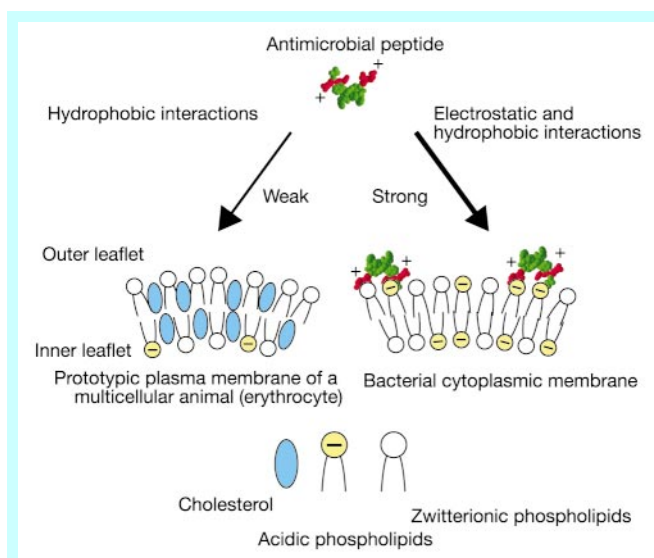


Figure 2 The membrane target of antimicrobial peptides of multicellular organisms and the basis of specificity. (Modified from ref. 21.)

leaflet of the bilayer, the surface exposed to the outer world, is heavily populated by lipids with negatively charged phospholipid headgroups. In contrast, the outer leaflet of the membranes of plants and animals is composed principally of lipids with no net charge; most of the lipids with negatively charged headgroups are segregated into the inner leaflet, facing the cytoplasm (Fig. 2) (reviewed in ref. 21). A model that explains the activity of most antimicrobial peptides is the Shai–Matsuzaki–Huang (SMH) model (refs 21–23; Fig. 3). The model proposes the interaction of

Table 1 Overview of antimicrobial peptides from plants and animals

Representative peptides		Origin	Tissue*
α -helical			
Cecropin A	KWKLKKIEKVGQNIIRDGIKAGPAVAVGQATQIAK _a	Silk moth	E, BC, H
Magainin 2	GIGKFLHSAKKFGKAFVGEIMNS	Frog	E
Pexiganan	GIGKFLKKAKKFGKAFVKILKK _a	Synthetic	
Dermaseptin 1	ALWKTMLKKGTLTALHAGKAALGAAADTISQGTQ	Frog	E
LL-37	LLGDFFRKSKKEKIGKEFKRIVQRIKDFLRNLVPRTES	Human	E, BC
Bufoin II	TRSSRAGLQFPVGRVHRLLRK	Vertebrate	E
One disulphide bond			
Bactenecin 1	RLCRIVIVRCR	Cow	BC
Thanatin	GSKKPVPIIYCNRRTGKQCQM	Insect	BC
Brevinin 1T	VNPIILGVLPKVLITKKC	<i>Rana</i> frogs	E
Ranalexin	FLGGLIKIVPAMICAVTKKC	<i>Rana</i> frogs	E
Ranateurin 1	SMLSVLKNLGKVLGFVACKINKQC	<i>Rana</i> frogs	E
Esculentin 1	GIFSKLGRKKIKNLLISGLKNVGEVGMVVRTGDIAGCKIKGEC	<i>Rana</i> frogs	E
Two disulphide bonds			
Tachyplesin	RWC ₁ FRVC ₂ YRGIC ₂ YRKC ₁ R _a	Horseshoe crab	BC
Androctonin	RSVC ₁ RQIKIC ₂ RRRGGC ₂ YYKC ₁ TNRPY	Scorpion	H
Protegrin 1	RGGRLC ₁ YC ₂ RRRFC ₂ VC ₁ VG _{Ra}	Pig	BC
Three disulphide bonds			
α -defensin (HNP3)	DC ₁ YC ₂ RIPAC ₃ IAGERRYGTC ₂ IYQGRWLWAF ₃ C ₁	Human	BC, E
β -defensin (TAP)	NPVSC ₁ VRNKGIC ₂ VPIRC ₃ PGSMKQIGTC ₂ VGRAVKC ₁ C ₃ RKK	Cow	E, BC
θ -defensin	GFC ₁ RC ₂ LC ₃ RRGVC ₃ RC ₂ IC ₁ TR	Monkey	BC
Defensin (sapecinA)	ATC ₁ DLLSGTGINSAC ₂ AAHC ₃ LLRGNRGGYC ₂ NGKAVC ₃ VC ₁ RN	Insect	E, BC, H
Thionin (crambin)	TTC ₁ C ₂ PISIVARSNFNC ₃ RIPGTPEAIC ₃ ATYTGC ₂ IIIPGATC ₁ PGDYAN	Plant	E
Four disulphide bonds			
Defensin	QKLC ₁ QRPSGTWSGVC ₂ GNNNAC ₃ KNQC ₄ IRLEKARHGSC ₂ NYVFAHC ₃ IC ₄ YFPC ₁	Radish	Seeds, E
Drosomycin	DC ₁ LSGRYKGPC ₂ AWDNETC ₃ RRVC ₄ KEEGRSSGHC ₂ SPSLKC ₃ WC ₄ EGC ₁	<i>Drosophila</i>	H
Hepcidin	DTHFPIC ₁ IFC ₂ C ₃ GC ₄ C ₁ HRSKC ₂ GMC ₃ C ₄ KT	Human	Liver
Linear, not α -helical			
Bac 5	RFRPPPIRRPPPIRPPFPFPPRPPPIRPPPIRPPFPPLGRPPF _a	Cow	BC
PR-39	RRRPPPPYLPFRPPPPFPPLPPRIPPFGPPFPFPFPF _a	Pig	BC
Indolicidin	ILPWKWPWWPWRR _a	Cow	BC
Apidaecin	GNNRPVYIPQPRPPHPRI	Honeybee	H
Pyrrhocoricin	VDKGSYLPPTPPRPPYINRN	Insect	H
Histatin 5	DSHAKRHHGYKRFHEKHHSHRGY	Human	Saliva

Cysteines paired in disulphide linkages are noted by common numerical subscripts. C-terminal amides are noted by *a*. In θ -defensin, the first and last residues are joined in a peptide bond.

* BC, blood cell; H, haemolymph; E, epithelial tissue.

the peptide with the membrane, followed by displacement of lipids, alteration of membrane structure, and in certain cases entry of the peptide into the interior of the target cell. The presence of cholesterol in the target membrane in general reduces the activity of antimicrobial peptides, due either to stabilization of the lipid bilayer or to interactions between cholesterol and the peptide²¹. Similarly, it is believed that increasing ionic strength, which in general reduces the activity of most antimicrobial peptides, does so in part by weakening the electrostatic charge interactions required for the initial interaction.

In general, peptides operating by the SMH mechanism kill microbes at micromolar concentrations. In contrast, the peptide nisin, a 14-amino-acid amphipathic molecule produced by *Lactococci*, operates at nanomolar concentrations. Nisin binds with high affinity to Lipid II, the fatty acyl proteoglycan anchor in the bacterial membrane, from which it subsequently diffuses into the surrounding membrane²⁴. Certain plant defensins use a similar strategy²⁵.

How do antimicrobial peptides actually kill microbes? Many hypotheses have been presented, which include: fatal depolarization of the normally energized bacterial membrane²⁶; the creation of physical holes that cause cellular contents to leak out²²; the activation of deadly processes such as induction of hydrolases that degrade the cell wall²⁷; the scrambling of the usual distribution of lipids between the leaflets of the bilayer, resulting in disturbance of membrane functions²¹; and the damaging of critical intracellular targets after internalization of the peptide, as suggested by the example of the peptide pyrrolicorin²⁸.

Resistance

Unlike conventional antibiotics such as penicillin, which microbes readily circumvent, acquisition of resistance by a sensitive microbial strain against antimicrobial peptides is surprisingly improbable. Resistant species of genera such as *Morganella* and *Serratia* express an outer membrane that lacks the appropriate density of acidic

lipids to provide peptide-binding sites. Other resistant species, such as *Porphyromonas gingivalis*, secrete digestive proteases that destroy peptides. Published studies of 'acquired resistance' against antimicrobial peptides, by and large, have identified genes that, when disrupted, make sensitive organisms more susceptible to a particular antimicrobial peptide; indeed, these genes usually appear to have a role in virulence.

A recent report of a survey of thousands of clinical isolates against the synthetic magainin analogue pexiganan illustrates the general picture that has emerged over the past decade regarding the issues of resistance²⁹. Bacterial species exhibit a wide range of susceptibilities, with some, such as anaerobes in the case of pexiganan, among the most sensitive. The basis for the different susceptibilities of bacterial and fungal species against particular peptides remains unexplained. Attempts at inducing pexiganan resistance in *Escherichia coli* and *Staphylococcus aureus* by chemical mutagenesis have been unsuccessful. As expected, no evidence of cross-resistance between pexiganan and any antibiotic in clinical use has been documented.

Gram-negative bacteria possess an outer membrane composed of lipopolysaccharide (LPS), which is held together by magnesium and calcium ions that bridge negatively charged phosphosugars. Addition of cationic peptides results in displacement of metal, damaging the outer membrane, and facilitates entry of additional molecules from the exterior (as reviewed in ref. 30). Peptides, having gained access to the periplasmic space, can now integrate into the cytoplasmic membrane. In many species of Gram-negative bacteria, the charge on the outer membrane is modulated by the PhoPQ regulon, a two-component system that uses a sensor (PhoQ) and an intracellular effector (PhoP)³¹. The PhoP/PhoQ regulon affects antimicrobial peptide sensitivity through modulation of the PmrA regulon, which controls a bank of genes that mediate decoration of the outer membrane with the positively charged moieties ethanolamine and 4-aminoarabinose³².

Why have microbes not been more successful in resisting the

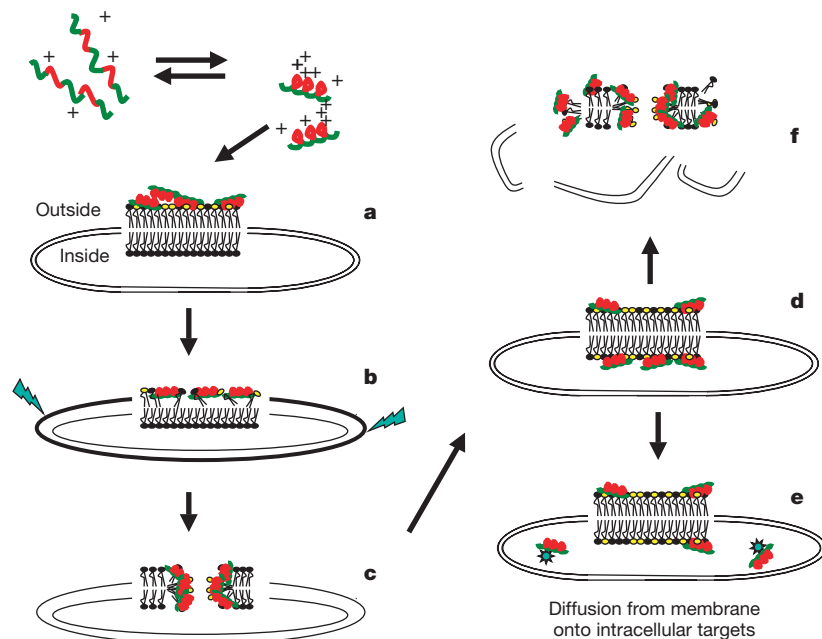


Figure 3 The Shai-Matsuzaki-Huang model of the mechanism of action of an antimicrobial peptide. An α -helical peptide is depicted. **a**, Carpeting of the outer leaflet with peptides. **b**, Integration of the peptide into the membrane and thinning of the outer leaflet. The surface area of the outer leaflet expands relative to the inner leaflet, resulting in strain within the bilayer (jagged arrows). **c**, Phase transition and

'wormhole' formation. Transient pores form at this stage. **d**, Transport of lipids and peptides into the inner leaflet. **e**, Diffusion of peptides onto intracellular targets (in some cases). **f**, Collapse of the membrane into fragments and physical disruption of the target cell's membrane. Lipids with yellow headgroups are acidic, or negatively charged. Lipids with black headgroups have no net charge.

activity of antimicrobial peptides, considering the time span over which such mechanisms could have evolved? Because the target of antimicrobial peptides is the bacterial membrane, a microbe would have to redesign its membrane, changing the composition and/or organization of its lipids, probably a 'costly' solution for most microbial species. Destruction of the antimicrobial peptide poses several problems. Most peptides are created from nondescript sequences of amino acids lacking unique epitopes that could serve as the recognition site of a protease required for selective destruction of the antibiotic in the presence of cellular protein constituents. In addition, multicellular organisms attack microbes with multiple peptides of different structural classes, hence destruction of one peptide might not suffice to ward off the lethal assault. Perhaps the virulence genes that bacteria now express which affect susceptibility towards antimicrobial peptides represent the best defences that most microbes can mount without suffering a loss of viability.

Regulation

Insects

After the discovery of inducible antimicrobial peptides in silk moths¹ and *Drosophila*³³, the corresponding genes were cloned and sequenced. The genes are expressed in blood cells, epithelia and the insects' fat body, an organ resembling the vertebrate liver that secretes proteins and peptides directly into the animals' haemolymph. The 5'-flanking regions harboured sequences that bound 'Rel' transcription factors, analogues of the nuclear factor κ B (NF κ B)-binding motifs of mammals^{34,35}. These putative regulatory regions had previously been implicated in early embryonic development of *Drosophila* (as reviewed in ref. 36). Examination of the intracellular pathway involved in the regulation of antimicrobial peptides revealed that much of the early embryonic circuit was co-opted for a defensive function. A recent view of this pathway (Fig. 4) shows it to involve initiation of the signal through proteolytic generation of the protein Spaetzle from its precursor; the presumed interaction of Spaetzle with a receptor called Toll; subsequent communication through a series of intracellular proteins resulting in the chemical modification of cytoplasmic Cactus, which is then

released from its physical union with the Dorsal-related protein Dif; the translocation of Dif into the nucleus, where it binds to a DNA sequence in the vicinity of the antimicrobial gene, activating transcription. Drosomycin, an important antifungal peptide in *Drosophila*, appears to be regulated by this circuit³⁶.

As predicted, mutant flies that have lost the function of certain genes within this pathway can no longer express Drosomycin after a fungal challenge and succumb to overwhelming fungal infection. Conversely, mutants that are created to overexpress Spaetzle, the Toll ligand, produce Drosomycin in the absence of fungal challenge. Mutants defective in the Toll–Drosomycin circuit can still express antibacterial peptides such as cecropin and dipterin³⁶. The residual circuit is called the *imd* pathway, regulated by a pathway that is activated by a Dif-related protein called relish³⁷. Like Dif, relish resides in the cytoplasm, but, in contrast, must be proteolytically cleaved by upstream events³⁷. The actual receptor that turns on the *imd* intracellular pathway has not yet been identified.

Certain general features of the antimicrobial systems of animals are illustrated by the *Drosophila* story. First, multiple 'hard-wired' circuits exist in an animal linking microbial assault to the expression of the genes of the corresponding antimicrobial peptide. Second, different circuits are wired to different banks of defensive peptides. The Toll pathway seems to have a role in fungal defence, whereas the *imd* pathway engages a different array of antimicrobial weapons to fend off bacteria, suggesting that these two pathways are intrinsically designed to respond to different physiological insults.

Microbes themselves have not been implicated as ligands of the insect Toll receptor. A recent report implicates a proteoglycan binding protein circulating in *Drosophila* haemolymph (seml) as the proximal receptor that initiates signals in the Toll pathway when the animal is invaded by Gram-positive bacteria (but not fungi)³⁸. It is not yet clear how the binding of this 'pattern recognition' receptor to proteoglycan turns on the proteolytic cascade that results in the generation of mature, processed Spaetzle from its precursor protein.

Vertebrates

Expression of the antimicrobial β -defensins TAP³⁹ in epithelial cells of the bovine respiratory tract and LAP in the tongue⁴⁰ are

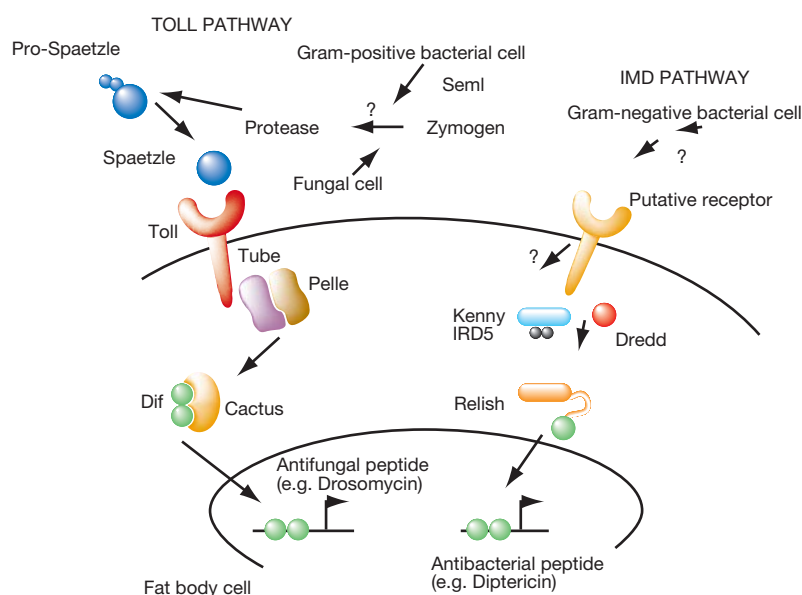


Figure 4 Signalling cascades that activate antimicrobial peptide genes in the fat body of *Drosophila*. The pathways have been deduced, by and large, through genetic analysis of immune function in *Drosophila*, and proteins encoded by genetic loci are

noted. The proteolytic system that results in the cleavage of pro-Spaetzle is not known. The existence of a receptor in the *Imd* circuit is depicted, but remains speculative.

stimulated by LPS, interleukin-1 β (IL-1 β) and tumour necrosis factor (TNF)⁴¹, and were upregulated *in vivo* in the setting of inflammation and after acute bacterial challenge. Reports followed describing induction of β -defensins in various epithelia including the human gastrointestinal⁴² and respiratory tracts^{43,44} and skin⁴⁵. Analysis of the 5'-flanking regions of several inducible genes of epithelial antimicrobial peptides from mammals^{46,47} and frogs (reviewed in ref. 4), like insects, revealed NF κ B-binding sites. Exposure to the appropriate stimulus resulted in increase in intracellular levels of NF κ B, its translocation into the nucleus, and activation of the corresponding antimicrobial gene.

The inducing activity of IL-1 β implicated a role for the IL-1 receptor (IL-1R) in the regulation of epithelial defensin production. The IL-1R is a close structural relative of Toll, and both respond to a protein ligand, which must be proteolytically processed. Furthermore, the intracellular pathways involved in IL-1 signalling converge on NF κ B like the corresponding Toll pathway in insects. In mammals, IL-1 can be liberated from monocytes, macrophages, dendritic cells, or from injured epithelial cells (reviewed in ref. 48).

The discovery that Toll and Toll-like receptors (TLRs) function in expression of genes for antimicrobial peptides in insects and mammals led to a search for additional human homologues, driven by the hypothesis that the innate immune system used specially tuned receptors to recognize specific and unique patterns of microbial chemical constituents that are presented when microbes attack a multicellular organism (reviewed in ref. 49). Subsequent studies have demonstrated that at least ten TLR genes exist in humans⁴⁹. Studies linking several of these TLRs with antimicrobial peptide gene expression in vertebrate systems have very recently been reported by several laboratories, but as yet are unpublished. The current model suggests that microbial products such as LPS bind directly to these TLRs, in some cases in association with specific binding proteins that might enhance specificity (such as the LPS-binding protein CD14), but the full story regarding ligand-receptor interactions as well the intracellular circuitry used to modulate antimicrobial peptide gene expression is still incomplete.

Plants

Like animals, plants express antimicrobial peptides such as defensins and thionins both constitutively and in response to microbial assault. The initial insult is transmitted by highly specific pattern-recognition receptors such as the nucleotide-binding site plus leucine-rich repeat with Toll and IL-1 receptor homology (TIR-ND-LRR) receptors, analogues of the mammalian TLRs; they in turn activate a relatively specific hard-wired response against the specific organisms to which they are tuned, a concept termed the 'gene-for-gene' defensive strategy (reviewed in ref. 50).

Roles in health and disease

Epithelial surfaces

Antimicrobial peptides are expressed by the epithelial cells within the barrier and delivered to the site by circulating white cells. This generalization is valid for all vertebrates studied, and probably every invertebrate, including insects⁵¹. In humans and other mammals, fully processed active peptides can be isolated from keratinized epithelial sheets of dry skin and tongue, where they probably act as epithelial 'preservatives'. In other sites in humans, such as the columnar epithelium of the airway, antimicrobial peptides are secreted into a micrometre-thick biofilm directly overlying the epithelium. The concentration and composition of electrolytes within this antimicrobial-rich unstirred layer are regulated by a variety of pumps and channels, which maintain conditions that maximize the antimicrobial activity of these defences. In cystic fibrosis, this homeostasis is disturbed by the genetic defect in the cystic fibrosis transmembrane-conductance regulator (CFTR); it has been suggested that, in this setting, antimicrobial peptide

defences fail to provide an effective barrier, resulting in bacterial overgrowth and chronic, tissue-destructive inflammation^{52,53}.

The skin of an animal is under constant microbial assault, and a recent study highlights the defensive contribution to this barrier provided by antimicrobial peptides. The mouse has only a single member of the cathelicidin gene family, called Cramp, which it expresses in both leukocytes and epithelial tissues. Knockout mutants of Cramp exhibit hypersusceptibility to group A *Streptococcus*, and succumb to destructive necrotic ulceration after inoculation of a dose of bacteria that causes only a mild self-limited reaction in wild-type mice. Similarly, mutants of group A streptococci selected for reduced susceptibility to Cramp (which, curiously, seem to grow more slowly than the wild type), exhibit an enhanced virulence when introduced into the dermis of wild-type mice⁵⁴.

The gastrointestinal tract of mammals is covered by a continuous sheet of epithelial cells that is folded into villus projections and crypts. Within the base of the crypts in humans and mice, where the stem cells of the GI tract can be found, are specialized, granular cells called Paneth cells. Both the enterocytes and the Paneth cells produce antimicrobial peptides (as reviewed in ref. 55). The enterocytes synthesize and secrete antimicrobial peptides both constitutively and on induction, and either secrete them onto their surface, as in the respiratory tract, or retain them in a cell-associated fashion in the superficial non-viable sheets of epithelium, as in the rectum. In contrast, after a microbial stimulus the Paneth cells at the base of the intestinal crypts secrete α -defensins into the cryptal well, at concentrations estimated at the level of grams per litre, which eventually flush into the gut lumen⁵⁶.

Both systems contribute to bowel health. In children and adults suffering from diarrhoea caused by *Shigella*, synthesis of the colonic enterocyte β -defensin HBD-1 and the cathelicidin LL37 is markedly depressed; expression recovers in time during resolution of the illness⁵⁷. Similarly, mice lacking the proteolytic enzyme required for processing cryptidins, which are the α -defensins of murine Paneth cells, and consequently lacking functional cryptidins, exhibit increased susceptibility to orally administered *Salmonella*⁵⁸.

A rather surprising pattern has been discovered operating in the gastric mucosa of humans, mice and other vertebrates¹². It appears that histone 2A is synthesized in excess of the needs required for DNA packaging in the gastric mucosal cell, and accumulates within cytoplasmic secretory granules. On secretion, the histone is processed by pepsin to the potent antimicrobial peptide buforin II, which remains adherent to the mucous biofilm coating the stomach surface, thus providing the stomach with a protective antimicrobial coat.

The healthy human is inhabited by a population of bacteria called 'commensals', which include organisms such as *Fusobacterium nucleatum* in the mouth and *Lactobacillus* species in the gut. These bacteria are relatively resistant to the action of endogenous antimicrobial peptides²⁹. Commensal bacteria are generally regarded as beneficial to the host. They can suppress pathogens by displacing them from a microbial niche or by secreting antimicrobial substances. Recent data suggest that commensals also provide protection by chronically stimulating epithelial surfaces to express antimicrobial peptides at levels that kill pathogen microbes¹⁵. In the gingival epithelium, *F. nucleatum* stimulates the inducible defensin HBD-2, whereas *P. gingivalis*, the anaerobe that destroys gum tissue, does not, behaving as a silent invader⁵⁹. Frogs that have been pharmacologically depleted of skin antimicrobial peptides will not re-accumulate skin antimicrobial peptides unless the animals are gradually exposed to bacteria in their environment; in their depleted state these frogs will succumb to overwhelming infection if suddenly exposed to otherwise innocuous microbes⁶⁰.

Sentinels of adaptive immunity

Antimicrobial peptides, released from circulating cells or induced in epithelia, can alert the adaptive immune system to trouble brewing.

Table 2 Commercial development of antimicrobial peptides of animal origin

Mode of use	Peptide	Company	Application	Stage
Topical	Pexiganan (MSI-78)	Magainin (Genaera)	Infected diabetic foot ulcers	Completed Phase III; not approved by FDA, pending additional studies
Topical	MBI-226	Micrologix	Catheter infection	Phase III
Topical	MBI-594	Micrologix	Acne	Phase II
Oral	Protegrin analogue (IB-367)	Intrabiotics	Mucositis	Phase III
Oral	Histatin analogue (P-113)	Demegen	Gingivitis	Phase II
Systemic	Heliomycin	Entomed	Antifungal	Preclinical
Systemic	Human lactoferricin	AM Pharma	Antibacterial	Preclinical
Systemic	BPI*	Xoma	Meningococcal meningitis	Phase III

* BPI, bactericidal permeability increasing protein.

The α -defensins of the human neutrophil directly attract human peripheral blood T cells that express CD4/CD45RA (naive) and CD8 antigens; the α -defensins also attract immature dendritic cells both *in vitro* and after injection under the skin of mammals (as reviewed in ref. 61). When administered simultaneously with antigens, these neutrophil defensins enhance antigen-specific immune responses. The cathelicidin LL37 attracts neutrophils along with monocytes and certain peripheral T cells. LL37 seems to act through a receptor for formyl peptide receptor-like 1 (FPRL1), a G-protein-coupled receptor that also recognizes ligands such as the bacterial formyl peptides⁶². LL37 is induced within epithelial cells of skin and lung in states of inflammation by some 10–50-fold, and would be expected to attract neutrophils, monocytes and T cells to the site of damage. The epithelial β -defensins, constitutively expressed HBD-1, and the inducible HBD-2 and HBD-3 are also chemoattractants. HBD-2 selectively attracts the memory subset of peripheral T cells (CD4⁺/CD45RO⁺), along with immature dendritic cells. The receptor used in this case is chemokine (C-C) receptor 6 (CCR6), which also recognizes the selective dendritic cell attractant macrophage inflammatory protein (MIP)-3 α ⁶¹. What is striking about the chemokine role ascribed to antimicrobial peptides is their relatively low affinity (10^{-6} to 10^{-5} M) compared with optimal concentrations observed for classical chemotactic factors (10^{-8} to 10^{-7} M). A cell relying solely on high-affinity ligands for directing movement might be stunned as it approaches the source of the chemokine. However, low-affinity interactions will still operate, and because of the presence of a large concentration gradient, cellular traffic will be directed with greater precision to specific sites of injury.

Applications

The growing problem of resistance to conventional antibiotics and the need for new antibiotics has stimulated interest in the development of antimicrobial peptides as human therapeutics. Most pharmaceutical effort has been devoted to the development of topically applied agents, such as the magainin analogue pexiganan⁶³, largely because of the relative safety of topical therapy and the uncertainty surrounding the long-term toxicology of any new class of drug administered systemically. The main hurdles that have impeded the development of antimicrobial peptides as systemic therapy include the unpublished experience that many of the naturally occurring peptides (such as magainin), although active *in vitro*, are effective in animal models of infection only at very high doses, often close to the toxic doses of the peptide⁶⁴, reflecting an unacceptable margin of safety. Antimicrobial peptides in pharmaceutical development are shown in Table 2.

Diverse applications have been demonstrated for antimicrobial peptides as anti-infective agents. The broad antimicrobial spectrum of antimicrobial peptides positions them for consideration as 'chemical condoms' to limit the spread of sexually transmitted diseases, including *Neisseria*, *Chlamydia*, human immunodeficiency virus (HIV) and *Herpes simplex* virus (HSV)⁶⁵. The affinity of antimicrobial peptides for microbial membranes has

encouraged their evaluation as imaging probes for bacterial and fungal infections⁶⁶. Antimicrobial peptides can enhance the potency of existing antibiotics *in vivo*, probably by facilitating access of antibiotics into the bacterial cell^{64,67}, a phenomenon previously recognized for the cationic peptide component of polymyxin.

Microbial colonization and growth on the surfaces of synthetic polymeric materials is a problem that complicates the use of medical devices such as intravenous catheters. One solution has been suggested by the successful demonstration that magainin peptides, covalently bound to insoluble polymeric beads, retain antimicrobial activity⁶⁸.

Introduction of antimicrobial genes into both plants and animals has been successful in transferring some benefit against disease. Agricultural uses have progressed most extensively, as demonstrated in tobacco⁶⁹ and potatoes⁷⁰. The possibility of alleviating the pulmonary bacterial infections associated with cystic fibrosis by transferring in a genetic construct capable of expressing super-physiological levels of LL37 has been demonstrated in an animal model⁷¹. The potential of re-engineering human macrophages to express β -defensins to enhance their efficacy against *Mycobacterium tuberculosis* has recently been proposed⁷².

Lastly, the discovery that the essential amino acid isoleucine can pharmacologically stimulate expression of β -defensin genes in isolated enteric cells suggests that new classes of therapeutics could be developed on the basis of their ability to turn on endogenous antimicrobial peptides⁷³.

Conclusion

The discovery of the widespread distribution of antimicrobial peptides over the past 20 years has provided insights into the innate defensive systems that permit multicellular organisms, including humans, to live in harmony with microbes. It is hard to imagine (especially after reading a classic immunology text) that most animals now alive, including insects and creatures like octopus and starfish, rely heavily on antimicrobial peptides for defence against microbes, and do so quite effectively without the help of lymphocytes, a thymus, or antibodies. If the story regarding diversity continues to unfold based on our current views, we might well discover that every species harbours a unique, specific collection of antimicrobial peptides, tuned to defend the organism against microorganisms that it will predictably encounter. Newly characterized molecules have inspired molecular designs for the creation of therapeutics, and will continue to do so as more are discovered, because these are based on antimicrobial strategies that have proven efficacious over millennia. Studies both in the laboratory and in the clinic confirm that emergence of resistance against antimicrobial peptides is less probable than observed for conventional antibiotics, and provides the impetus to develop antimicrobial peptides, both natural and laboratory conceived, into therapeutically useful agents.

Note added in proof: Bacterial lipoprotein has been shown to induce the expression of HBD-2 in human epithelial cell lines through a TLR-2/NF κ B-dependent pathway⁷⁴. □

1. Steiner, H., Hultmark, D., Engstrom, A., Bennich, H. & Boman, H. G. Sequence and specificity of two antibacterial proteins involved in insect immunity. *Nature* **292**, 246–268 (1981).
2. Zasloff, M. Magainins, a class of antimicrobial peptides from *Xenopus* skin: isolation, characterization of two active forms, and partial cDNA sequence of a precursor. *Proc. Natl Acad. Sci. USA* **84**, 5449–5453 (1987).
3. Bechinger, B., Zasloff, M. & Opella, S. J. Structure and orientation of the antibiotic peptide magainin in membranes by solid-state nuclear magnetic resonance spectroscopy. *Protein Sci.* **2**, 2077–2084 (1993).
4. Simmaco, M., Mignogna, G. & Barra, D. Antimicrobial peptides from amphibian skin: what do they tell us? *Biopolymers* **47**, 435–450 (1998).
5. Romeo, D., Skerlavaj, B., Bolognesi, M. & Gennaro, R. Structure and bactericidal activity of an antibiotic dodecapeptide purified from bovine neutrophils. *J. Biol. Chem.* **263**, 9573–9575 (1988).
6. Selsted, M. E., Harwig, S. S., Ganz, T., Schilling, J. W. & Lehrer, R. I. Primary structure of three human neutrophil defensins. *J. Clin. Invest.* **76**, 1436–1439 (1985).
7. Selsted, M. E. *et al.* Indolicidin, a novel bactericidal tridecapeptide amide from neutrophils. *J. Biol. Chem.* **267**, 4292–4295 (1992).
8. Agerberth, B. *et al.* Amino acid sequence of PR39. Isolation from pig intestine of a new member of the family of proline arginine rich antibacterial peptides. *Eur. J. Biochem.* **202**, 849–854 (1991).
9. Bulet, P. *et al.* A novel inducible antibacterial peptide of *Drosophila* carries an O-glycosylated substitution. *J. Biol. Chem.* **268**, 14893–14897 (1993).
10. Shinmar, A. E. *et al.* in *Peptides: Chemistry and Biology. Proc. 14th Am. Peptide Symp.* (eds Kaumaya, P. & Hodges, R.) 189–191 (Mayflower Scientific, Leiden, 1996).
11. Tang, Y. Q. *et al.* A cyclic antimicrobial peptide produced in primate leukocytes by the ligation of two truncated α -defensins. *Science* **286**, 498–502 (1999).
12. Kim, H. S. *et al.* Pepsin mediated processing of the cytoplasmic histone 2A to the strong antimicrobial peptide Buforin I. *J. Immunol.* **165**, 3268–3274 (2000).
13. Ulvatne, H. & Vorland, L. H. Bactericidal kinetics of three lactoferricins against *S. aureus* and *E. coli*. *Scand. J. Infect. Dis.* **33**, 507–511 (2001).
14. Zanetti, M., Gennaro, R., Scochi, M. & Skerlavaj, B. Structure and biology of cathelicins. *Adv. Exp. Med. Biol.* **479**, 203–218 (2000).
15. Boman, H. G. Innate immunity and the normal microflora. *Immunol. Rev.* **173**, 5–16 (2000).
16. Maloy, W. L. & Kari, U. P. Structure-activity studies on magainins and other host defense peptides. *Biopolymers* **37**, 105–122 (1995).
17. Fernandez-Lopez, S. *et al.* Antibacterial agents based on the cyclic α -peptide architecture. *Nature* **412**, 452–455 (2001).
18. Oren, Z. & Shai, Y. Cyclization of a cytolytic amphipathic α -helical peptide and its diastereomer: effect on structure, interaction with model membranes, and biological function. *Biochemistry* **39**, 6103–6114 (2000).
19. Porter, E. A., Wang, X., Lee, H. S., Weisblum, B. & Gellman, S. H. Non-haemolytic β -amino acid oligomers. *Nature* **404**, 565 (2000).
20. Hamuro, Y., Schneider, J. P. & DeGrado, W. F. De novo design of antibacterial β -peptides. *J. Am. Chem. Soc.* **121**, 12200–12201 (1999).
21. Matsuzaki, K. Why and how are peptide–lipid interactions utilized for self-defense? Magainins and tachyplesins as archetypes. *Biochim. Biophys. Acta* **1462**, 1–10 (1999).
22. Yang, L., Weiss, T. M., Lehrer, R. I. & Huang, H. W. Crystallization of antimicrobial pores in membranes: magainin and protegrin. *Biophys. J.* **79**, 2002–2009 (2000).
23. Shai, Y. Mechanism of the binding, insertion and destabilization of phospholipid bilayer membranes by α -helical antimicrobial and cell non-selective membrane-lytic peptides. *Biochim. Biophys. Acta* **1462**, 55–70 (1999).
24. Brotz, H. *et al.* Role of lipid-bound peptidoglycan precursors in the formation of pores by nisin, epidermin and other lantibiotics. *Mol. Microbiol.* **30**, 317–327 (1998).
25. Thevissen, K. *et al.* A gene encoding a sphingolipid biosynthesis enzyme determines the sensitivity of *Saccharomyces cerevisiae* to an antifungal plant defensin from dahlia (*Dahlia merckii*). *Proc. Natl Acad. Sci. USA* **97**, 9531–9536 (2000).
26. Westerhoff, H. V., Juretic, D., Hendler, R. W. & Zasloff, M. Magainins and the disruption of membrane-linked free-energy transduction. *Proc. Natl Acad. Sci. USA* **86**, 6597–6601 (1989).
27. Bierbaum, G. & Sahl, H.-G. Induction of autolysis of *Staphylococci* by the basic peptide antibiotics pep5 and nisin and their influence on the activity of autolytic enzymes. *Arch. Microbiol.* **141**, 249–254 (1985).
28. Kragol, G. *et al.* The antibacterial peptide pyrrolicorin inhibits the ATPase actions of DnaK and prevents chaperone-assisted protein folding. *Biochemistry* **40**, 3016–3026 (2001).
29. Ge, Y. *et al.* In vitro antibacterial properties of pexiganan, an analog of magainin. *Antimicrob. Agents Chemother.* **43**, 782–788 (1999).
30. Hancock, R. E., Falla, T. & Brown, M. Cationic bactericidal peptides. *Adv. Microbial Phys.* **37**, 135–175 (1995).
31. Groisman, E. A. The ins and outs of virulence gene expression: Mg^{2+} as a regulatory signal. *Bioessays* **20**, 96–101 (1998).
32. Gunn, J. S., Ryan, S. S., Van Velkinburgh, J. C., Ernst, R. K. & Miller, S. I. Genetic and functional analysis of a PmrA–PmrB-regulated locus necessary for lipopolysaccharide modification, antimicrobial peptide resistance, and oral virulence of *Salmonella enterica* serovar typhimurium. *Infect. Immun.* **68**, 6139–6146 (2000).
33. Kylsten, P., Samakovlis, C. & Hultmark, D. The cecropin locus in *Drosophila*: a compact gene cluster involved in response to infection. *EMBO J.* **9**, 217–224 (1990).
34. Engström, Y. *et al.* κ B like motifs regulate the induction of immune genes in *Drosophila*. *J. Mol. Biol.* **232**, 327–333 (1993).
35. Kappler, C. *et al.* Insect immunity. Two 17 bp repeats nest a κ B related sequence confer inducibility to the dipterin gene and bind a polypeptide in bacteria challenged *Drosophila*. *EMBO J.* **12**, 1561–1568 (1993).
36. Imler, J. L. & Hoffmann, J. A. Toll receptors in innate immunity. *Trends Cell Biol.* **11**, 304–311 (2001).
37. Stoven, S., Ando, I., Kadalayil, L., Engstrom, Y. & Hultmark, D. Activation of the *Drosophila* NF- κ B factor relish by rapid endoproteolytic cleavage. *EMBO Rep.* **1**, 347–352 (2000).
38. Michel, T., Reichhart, J.-M., Hoffmann, J. & Royet, J. *Drosophila* Toll is activated by Gram-positive bacteria via a circulating peptidoglycan recognition protein. *Nature* **414**, 756–759 (2001).
39. Diamond, G. *et al.* Tracheal antimicrobial peptide, a cysteine-rich peptide from mammalian tracheal mucosa: peptide isolation and cloning of a cDNA. *Proc. Natl Acad. Sci. USA* **88**, 3952–3956 (1994).
40. Schonwetter, B. S., Stolzenberg, E. D. & Zasloff, M. Epithelial antibiotics induced at sites of inflammation. *Science* **267**, 1645–1648 (1995).
41. Russell, J. P., Diamond, G., Tarver, A. P., Scanlin, T. F. & Bevins, C. L. Coordinate induction of two antibiotic genes in tracheal epithelial cells exposed to the inflammatory mediators LPS and tumor necrosis factor α . *Infect. Immun.* **64**, 1565–1568 (1996).
42. O’Neil, D. A. *et al.* Expression and regulation of the human β -defensins HBD-1 and HBD-2 in intestinal epithelium. *J. Immunol.* **163**, 6718–6724 (1999).
43. Singh, P. K. *et al.* Production of β -defensins by human airway epithelia. *Proc. Natl Acad. Sci. USA* **95**, 14961–14966 (1998).
44. Harder, J. *et al.* Mucoid *Pseudomonas aeruginosa*, TNF- α , and IL-1 β , but not IL-6, induce human β defensin-2 in respiratory epithelia. *Am. J. Respir. Cell Mol. Biol.* **22**, 714–721 (2000).
45. Harder, J., Bartels, J., Christophers, E. & Schroder, J. M. A peptide antibiotic from human skin. *Nature* **387**, 861 (1997).
46. Tarver, A. P. *et al.* Enteric β -defensin: molecular cloning and characterization of a gene with inducible intestinal epithelial cell expression associated with *Cryptosporidium parvum* infection. *Infect. Immun.* **66**, 1045–1056 (1998).
47. Liu, L. *et al.* Structure and mapping of the human β -defensin 2 gene and its expression at sites of inflammation. *Gene* **222**, 237–244 (1998).
48. Murphy, J. E., Robert, C. & Kupper, T. S. Interleukin-1 and cutaneous inflammation: a crucial link between innate and acquired immunity. *J. Invest. Dermatol.* **114**, 602–608 (2000).
49. Kimbrell, D. A. & Beutler, B. The evolution and genetics of innate immunity. *Nature Rev. Genet.* **2**, 256–267 (2001).
50. Dangl, J. L. & Jones, J. D. Plant pathogens and integrated defence responses to infection. *Nature* **411**, 826–833 (2001).
51. Brey, P. T. *et al.* Role of the integument in insect immunity: epicuticular abrasion and induction of cecropin synthesis in cuticular epithelial cells. *Proc. Natl Acad. Sci. USA* **90**, 6275–6279 (1993).
52. Smith, J. J., Travis, S. M., Greenberg, E. P. & Welsh, M. J. Cystic fibrosis airway epithelia fail to kill bacteria because of abnormal airway surface fluid. *Cell* **85**, 229–236 (1996).
53. Goldman, M. J. *et al.* Human β -defensin-1 is a salt-sensitive antibiotic in lung that is inactivated in cystic fibrosis. *Cell* **88**, 553–560 (1997).
54. Nizet, V. *et al.* An antimicrobial peptide protects mice from invasive bacterial infection. *Nature* **414**, 454–457 (2001).
55. Ouellette, A. J. & Bevins, C. L. Paneth cell defensins and innate immunity of the small bowel. *Inflamm. Bowel Dis.* **7**, 43–50 (2001).
56. Ayabe, T. *et al.* Secretion of microbicidal α -defensins by intestinal Paneth cells in response to bacteria. *Nature Immunol.* **1**, 113–118 (2000).
57. Islam, D. *et al.* Down-regulation of bactericidal peptides in enteric infections: A novel immune escape mechanism with bacterial DNA as a potential regulator. *Nature Med.* **7**, 180–185 (2001).
58. Wilson, C. L. *et al.* Regulation of intestinal α -defensin activation by the metalloproteinase matrilysin in innate host defense. *Science* **286**, 113–117 (1999).
59. Krisanaprakornkit, S. *et al.* Inducible expression of human β -defensin 2 by *Fusobacterium nucleatum* in oral epithelial cells: multiple signaling pathways and role of commensal bacteria in innate immunity and the epithelial barrier. *Infect. Immun.* **68**, 2907–2915 (2000).
60. Mangoni, M. L., Miele, R., Renda, T. G., Barra, D. & Simmaco, M. The synthesis of antimicrobial peptides in the skin of *Rana esculenta* is stimulated by microorganisms. *FASEB J.* **8**, 1431–1432 (2001).
61. Chertov, O., Yang, D., Howard, O. M. & Oppenheim, J. J. Leukocyte granule proteins mobilize innate host defenses and adaptive immune responses. *Immunol. Rev.* **177**, 68–78 (2000).
62. Yang, D. *et al.* LL-37, the neutrophil granule and epithelial cell-derived cathelicidin, utilizes formyl peptide receptor-like 1 (FRPL1) as a receptor to chemoattract human peripheral blood neutrophils, monocytes and T cells. *J. Exp. Med.* **192**, 1069–1074 (2000).
63. Zasloff, M. in *From Development of Novel Antimicrobial Agents: Emerging Strategies* (ed Lohner, K.) 261–270 (Horizon Scientific, Wymondham, UK, 2001).
64. Darveau, R. P. *et al.* β -lactam antibiotics potentiate Magainin 2 antimicrobial activity *in vitro* and *in vivo*. *Antimicrob. Agents Chemother.* **35**, 1153–1159 (1991).
65. Yasin, B. *et al.* Evaluation of the inactivation of infectious *Herpes simplex* virus by host-defense peptides. *Eur. J. Clin. Microbiol. Infect. Dis.* **19**, 187–194 (2000).
66. Welling, M. M., Paulusma-Annema, A., Balter, H. S., Pauwels, E. K. & Nibbering, P. H. Technetium-99m labeled antimicrobial peptides discriminate between bacterial infections and sterile inflammations. *Eur. J. Nucl. Med.* **27**, 292–301 (2000).
67. Giacometti, A., Cirioni, O., Barchiesi, F. & Scalise, G. In-vitro activity and killing effect of polycationic peptides on methicillin-resistant *Staphylococcus aureus* and interactions with clinically used antibiotics. *Diagn. Microbiol. Infect. Dis.* **38**, 115–118 (2000).
68. Haynie, S. L., Crum, G. A. & Doebe, B. A. Antimicrobial activities of amphiphilic peptides covalently bonded to a water-insoluble resin. *Antimicrob. Agents Chemother.* **39**, 301–307 (1995).
69. De Gray, G., Rajasekaran, K., Smith, F., Sanford, J. & Daniell, H. Expression of an antimicrobial peptide via the chloroplast genome to control phytopathogenic bacteria and fungi. *Plant Physiol.* **127**, 852–862 (2001).
70. Osusky, M. *et al.* Transgenic plants expressing cationic peptide chimeras exhibit broad-spectrum resistance to phytopathogens. *Nature Biotechnol.* **18**, 1162–1166 (2000).
71. Bals, R., Weiner, D. J., Mosconi, A. D., Meegalla, R. L. & Wilson, J. M. Augmentation of innate host defense by expression of a cathelicidin antimicrobial peptide. *Infect. Immun.* **67**, 6084–6089 (1999).
72. Kisch, K. O., Heifets, L., Higgins, M. & Diamond, G. Antimycobacterial agent based on mRNA encoding human β -Defensin 2 enables primary macrophages to restrict growth of *Mycobacterium tuberculosis*. *Infect. Immun.* **69**, 2692–2699 (2001).
73. Fehlbaum, P., Rao, M., Zasloff, M. & Anderson, G. M. An essential amino acid induces epithelial β -defensin expression. *Proc. Natl Acad. Sci. USA* **97**, 12723–12728 (2000).
74. Birchler, T. *et al.* Human Toll-like receptor 2 mediates induction of the antimicrobial peptide human β defensin 2 in response to bacterial lipoprotein. *Eur. J. Immunol.* **31**, 3131–3137 (2001).

Acknowledgements

I thank A. Nelson for his critical reading of the manuscript. I apologize to my many colleagues whose work I have not cited, owing to constraints on the length of this review.

Correspondence and requests for material should be addressed to the author (e-mail: maz5@georgetown.edu).

Structural basis for the activation of anthrax adenylyl cyclase exotoxin by calmodulin

Chester L. Drum^{*†‡}, Shui-Zhong Yan^{*}, Joel Bard[†], Yue-Quan Shen^{*}, Dan Lu^{*}, Sandriyana Soelaiman^{*}, Zenon Grabarek[†], Andrew Bohm^{†§||} & Wei-Jen Tang^{*‡||}

^{*}Ben-May Institute for Cancer Research, The University of Chicago, 924 East 57th Street; and [‡]Committee on Neurobiology, The University of Chicago, 947 East 58th Street, Chicago, Illinois 60637, USA

[†]Boston Biomedical Research Institute, 64 Grove Street, Watertown, Massachusetts 02472, USA

[§]Tufts University School of Medicine, 136 Harrison Avenue, Boston, Massachusetts 02111, USA

^{||}These authors contributed equally to this work

Oedema factor, a calmodulin-activated adenylyl cyclase, is important in the pathogenesis of anthrax. Here we report the X-ray structures of oedema factor with and without bound calmodulin. Oedema factor shares no significant structural homology with mammalian adenylyl cyclases or other proteins. In the active site, 3'-deoxy-ATP and a single metal ion are well positioned for catalysis with histidine 351 as the catalytic base. This mechanism differs from the mechanism of two-metal-ion catalysis proposed for mammalian adenylyl cyclases. Four discrete regions of oedema factor form a surface that recognizes an extended conformation of calmodulin, which is very different from the collapsed conformation observed in other structures of calmodulin bound to effector peptides. On calmodulin binding, an oedema factor helical domain of relative molecular mass 15,000 undergoes a 15 Å translation and a 30° rotation away from the oedema factor catalytic core, which stabilizes a disordered loop and leads to enzyme activation. These allosteric changes provide the first molecular details of how calmodulin modulates one of its targets.

Many infectious organisms increase the concentration of cyclic AMP in infected host cells, thereby disrupting intracellular signalling pathways. Whereas some bacteria achieve this through the chemical modification of heterotrimeric G proteins, three bacteria, *Bacillus anthracis* (anthrax), *Bordetella pertussis* (whooping cough) and *Pseudomonas aeruginosa* (various nosocomial infections), increase cAMP concentrations by producing exotoxins with adenylyl cyclase activity^{1–3}.

Oedema factor (EF), the exotoxin from *B. anthracis*, is transported into host cells by an anthrax-derived transporter, protective antigen. In conjunction with lethal factor, another transported anthrax toxin that proteolytically inactivates mitogen-activated protein kinase kinase⁴, EF contributes significantly to both cutaneous and systemic anthrax⁵. EF is activated by calmodulin (CaM), which is naturally present in host cells⁶. The carboxy-terminal portion of EF (residues 291–800; relative molecular mass 58,000; *M_r* 58K) is sufficient to reproduce the CaM-dependent activation and catalytic rate of the intact toxin, which is 1,000-fold higher than that of mammalian CaM-activated adenylyl cyclase^{7,8}.

The prototypic Ca²⁺ ion sensor, CaM (*M_r* 16.5K) is present in all eukaryotes and modulates many intracellular processes including signal transduction, gene transcription, ion conductivities, vesicular fusion and cytoskeleton functions^{9–11}. CaM transduces intracellular Ca²⁺ signals through two globular domains connected by a flexible, central α-helix⁶. Each globular domain comprises two helix–loop–helix, EF-hand Ca²⁺-binding motifs¹². Calcium binding shifts each of these domains from a mainly hydrophilic closed state to an open conformation that exposes a large, hydrophobic binding pocket^{13,14}.

The structure of CaM has been solved in complex with several biochemically inactive effector fragments^{15–18}. Here we report the crystal structures of the catalytic portion of EF alone, EF in complex with CaM, and EF in complex with both CaM and a non-cyclizable nucleotide analogue, 3'-deoxy-ATP (3'dATP). A comparison of these three structures provides a detailed description of the catalytic

mechanism and the CaM-induced activation of EF.

Structure of EF with and without bound CaM

The catalytic portion of EF comprises three globular domains (Figs 1 and 2). The active site lies at the interface of two domains, C_A (amino acids 294–349 and 490–622) and C_B (350–489), which together constitute the catalytic core. A third, helical domain (660–800) is connected to C_A by a linker (623–659). The structure of EF alone differs significantly from that of EF–CaM, but there are essentially no differences between the structures of EF–CaM with and without nucleotide (Fig. 1). Thus, the differences between EF alone and EF–CaM are almost entirely induced by CaM.

In the structure of EF alone, the interface between the helical domain and the remainder of the protein buries 16 hydrophobic residues and 3,600 Å² of surface area. None of these interactions are preserved in the EF–CaM complex because CaM inserts itself between C_A and the helical domain. In doing so, CaM occupies much of the same volume occupied by the helical domain in the structure of EF alone. To accommodate CaM the helical domain, acting as a unit, moves 15 Å and rotates 30° such that C_A, the linker and the helical domain form a large clamp that almost completely encircles CaM. There are no contacts between C_A and the helical domain in the EF–CaM structures, but residues from these domains approaching from either side of CaM come within 5.6 Å of one another at the tip of the clamp (Fig. 3). The EF–CaM interface buries 5,900 Å² of solvent-accessible surface area between the two proteins.

In addition to the helical domain movement, three segments designated as switch A, B and C undergo large conformational changes in response to CaM binding, leaving the remaining residues relatively unchanged (Figs 1 and 4). Switch A and C are the predominant regions that form the interface with the helical domain in the EF-alone structure (Fig. 2). Switch A (residues 502–551) includes amino acids that bind CaM and nucleotide. In the EF-alone structure, this loop is nestled on the side of domain C_A

and its tip is 20 Å away from its active state position. Switch C (630–659) is composed of residues from the linker that connects C_A to the helical domain. The tip of the switch C loop swings 33 Å from its position near the active site in EF–CaM to the opposite side of the enzyme in EF alone. In switch C, two β -strands and a connecting loop (642–652) are converted to a helix in the EF–CaM structure (648–653). Switch B (578–591), which lies between switch C and the nucleotide substrate, becomes disordered when CaM is not present. This disorder cannot be attributed to the lack of substrate in the EF-alone structure, because switch B is clearly visible in the structure of substrate-free EF–CaM (Fig. 4).

Geometry of the active site and catalytic mechanism

The non-cyclizable substrate analogue 3'dATP is clearly visible in simulated annealing omit maps of the EF–CaM–3'dATP structure (see Supplementary Information). Residues from six segments of the protein form the substrate-binding pocket (Fig. 5a). These residues form a network of hydrogen bonds and ionic interactions that encircle the triphosphate and ribose moieties completely, forming an enclosed cavity with openings at both ends (Fig. 5a).

The most notable phosphate interactions are made by Lys 346, which contacts oxygens from all three phosphates; by Arg 329, which interacts with the β -phosphate; and by Lys 372 and Ser 354, which interact with the γ -phosphate (Fig. 5a). Phe 586 and Leu 348 lie above and below the plane of the ribose, and the side chain of Asn 583 forms a hydrogen bond with O4' of the ribose ring. The

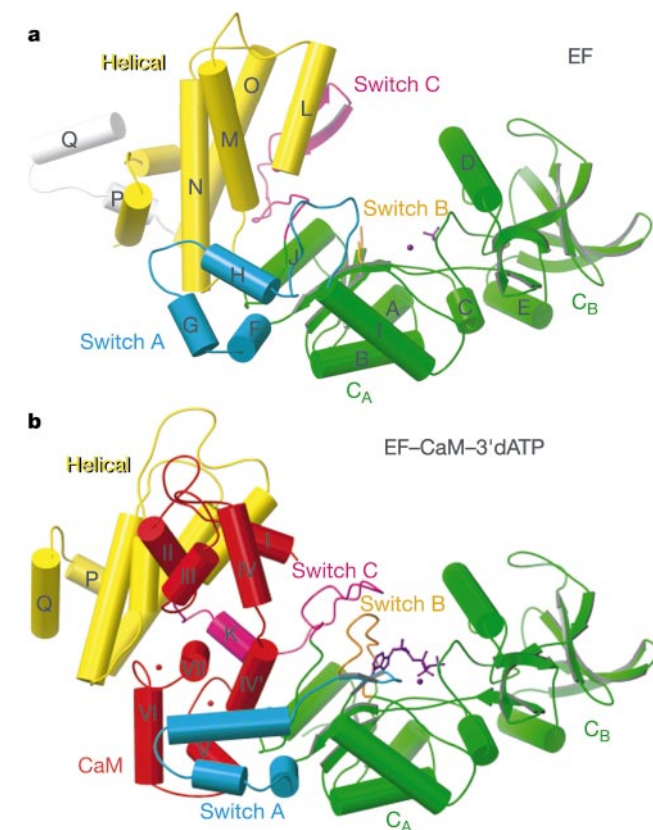


Figure 1 Secondary structures of EF. **a**, EF alone; **b**, The EF–CaM–3'-deoxy-ATP complex. CaM (red) is held on one side by a fingerlike projection of C_A (green), and on the other side by switch C (magenta) and the helical domain (yellow). Switch A and switch B are in cyan and orange, respectively. The 3'-deoxy-ATP analogue (purple) and metal (purple) are bound between the C_A and C_B domains (green), which form the active site. In the EF-alone structure, the last two helices, P and Q (amino acids 771–799, white), which are not part of the CaM interface or the active site, are swapped between two EF molecules. This swap is unlikely to be physiologically relevant because EF is a monomer on the basis of its gel-filtration properties.

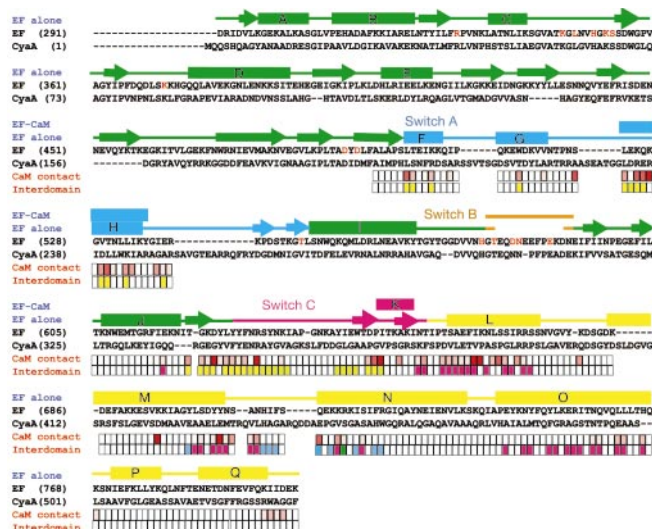


Figure 2 Sequence alignment of EF and CyaA. Residues that contact the nucleotide substrate are highlighted in red. The secondary structure assignment of EF alone is shown above the sequences, with the structures of the three switches of EF–CaM indicated on top. Molecular contacts are shown as coloured boxes below the sequences. CaM contact residues of EF are coloured on the basis of their CaM contact area in five shades of red representing increments of 20 Å². The lightest coloured boxes indicate less than 20 Å² of contact area, the darkest coloured boxes indicate that more than 80 Å² are buried by the CaM interaction. Interdomain indicates residues for the interdomain interaction between the helical domain and the remainder of EF in the EF-alone structure. Residues are coloured on the basis of their contact region: switch A, cyan; switch C, magenta; C_A region, green; and helical domain, yellow.

adenosine ring, although solvent accessible, contacts the backbone of Asp 582 and Asn 583, and the N6 nitrogen is within hydrogen-bonding distance of the main-chain carbonyls of Thr 579 and Thr 548. Mutational analysis of many of these residues supports their importance in catalysis (ref. 7 and Table 1).

At the base of the cleft formed by domains C_A and C_B , a single metal interacts with the oxygens of the α - and β -phosphates of 3'dATP. This metal is coordinated by two conserved aspartates, Asp 491 and Asp 493. Mutation of either aspartic acid in EF or of the homologous residues in CyaA and ExoY (exotoxins from *B. pertussis*

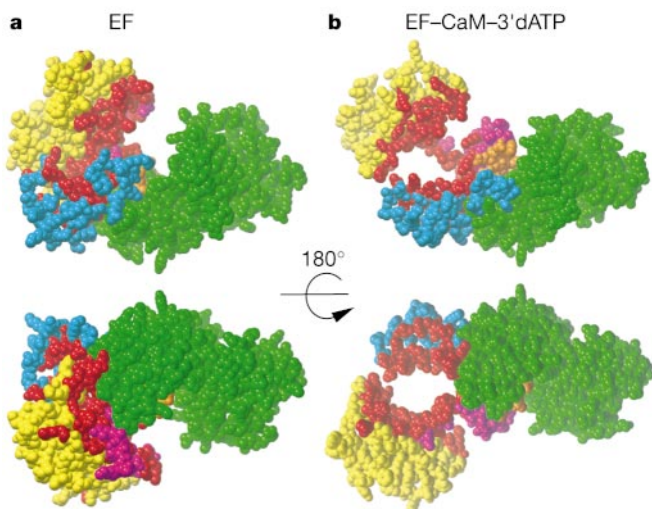


Figure 3 CPK representation of EF structures. **a**, EF alone; **b**, CaM–EF. CaM-contacting residues are in red, and the atoms for switch A, B and C are in cyan, orange and yellow, respectively.

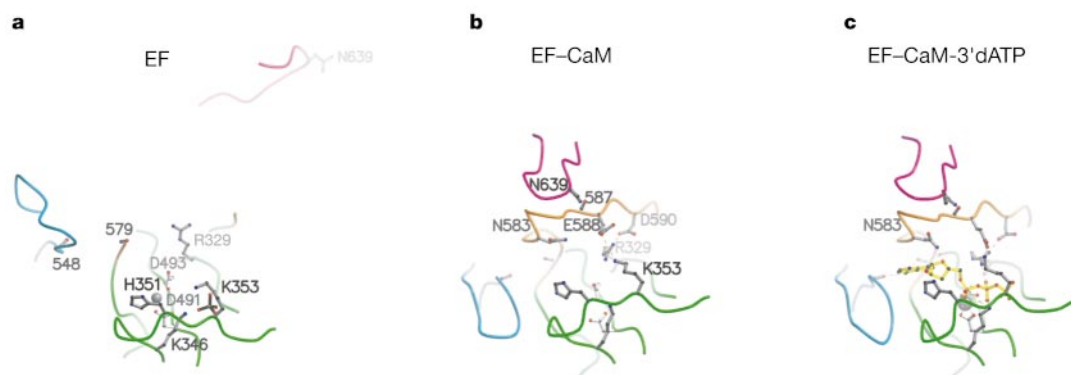


Figure 4 Switch movement results in catalytic activation. This view looks down on the molecule from above, relative to the view shown in Fig. 1. **a**, Ball-and-stick representation of the active site of the EF-alone structure; **b**, EF–CaM complex without nucleotide; **c**, EF–CaM complex with 3'-deoxy-ATP. SO_4 in the EF-alone structure and 3'dATP in

EF–CaM–3'dATP complex are in black and yellow, respectively, and metal ions are in grey. Residues from switch A (540–551), switch B (578–592), switch C (634–642) and $\text{C}_\alpha/\text{C}_\beta$ outside the switch regions are in cyan, orange, magenta and green, respectively.

and *P. aeruginosa*, respectively) results in enzyme inactivation (refs 3, 7, 19; and Fig. 2). In our highest resolution data set for EF–CaM–3'dATP, this metal is Yb^{3+} , a crystallization additive that only slightly reduces catalytic activity, rather than Mg^{2+} , the presumed physiological ligand (see Supplementary Information). A single metal ion (presumably Ni^{2+}) is also observed at this site in the structure of EF alone, which was crystallized in the presence of 220 mM NiSO_4 (Fig. 4a). As no metal ion was found in the catalytic site of EF–CaM in the absence of ATP analogue, it seems that the substrate may be required to bind the catalytic metal (Fig. 4b).

Histidine 351 lies on the opposite side of the ribose from the metal and is well positioned to interact with the reactive 3' hydroxyl when the ribose is in the 3'-endo conformation observed in the crystal (Fig. 5a). The homologous histidine in CyaA, His 63, acts as a catalytic base²⁰. Although the 3' hydroxyl is absent from the nucleotide analogue used in this study, the expected position of this atom is generally consistent with that required for in-line nucleophilic attack on the α -phosphate. In this reaction scheme, the metal and Lys 346 would both activate the α -phosphate and

stabilize the developing charge on the oxygen bridging the α - and β -phosphates (Fig. 5b). This catalytic mechanism differs from that proposed for mammalian adenylyl cyclases (mAC), which are thought to function by two-metal-ion catalysis²¹. Although His 577 seems to be positioned such that it could coordinate a second metal ion in conjunction with Asp 493, mutational analysis argues against such a role (Table 1). His 577Asn, a mutant with 180-fold reduced activity relative to the wild-type enzyme, has a 250-fold higher catalytic activity than His 577Asp, a mutation that would favour the binding of second metal ion.

Comparison of EF and mAC

All known adenylyl cyclases can be classified into four genetically unrelated groups. Although the catalytic core structures of mAC and expressed side-associated gene (ESAC) from *Trypanosoma brucei* (class III, the family of adenylyl cyclases from bacteria to mammals²²) have been determined crystallographically^{23,24}, EF represents the first structurally characterized member of the adenylyl cyclase toxin family (class II). These two classes show no obvious structural homology (Fig. 5d). Whereas the catalytic portion of the

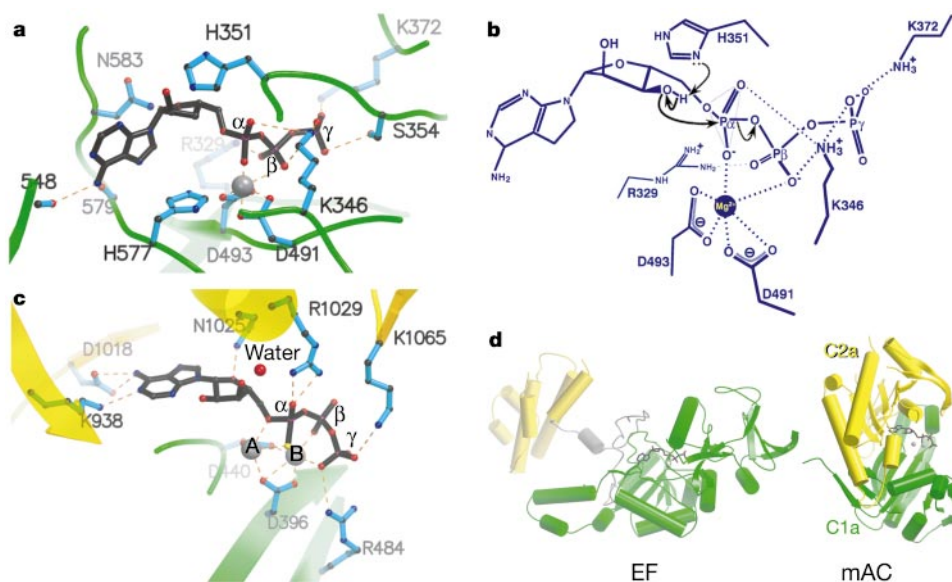


Figure 5 The active site of EF and its comparison with mAC. **a**, Ball-and-stick representation of the EF active site. **b**, Proposed mechanism of catalysis of EF. For clarity, only residues that are directly involved in 3' O' to P_α nucleophilic attack are shown. **c**, Ball-and-stick representation of the mAC active site. **d**, Secondary structures of the EF

catalytic core and the mAC catalytic core with the same view as in **a** and **c**. The nucleotide (3'dATP in EF–CaM and ATP- αS in mAC) and the metal ion are in black and grey, respectively. The O4', C4' and P_α of ATP- αS are shown in a similar position to those of 3'dATP in **a** for comparison. C1a and C2a of mAC are in green and yellow, respectively.

mammalian enzymes comprises two similarly folded domains (C1a and C2a) with rough two-fold symmetry, EF, which functions as a soluble monomer, has no such symmetry.

Superimposing the ATP- α S nucleotide from mAC²³ with the 3'dATP nucleotide seen in EF–CaM indicates, however, that these two families share amino-acid similarities in substrate binding and catalysis (Fig. 5a, c). Both enzymes have an asparagine that forms a hydrogen bond to the O4' of the ribose ring. Both enzymes also have a cluster of positively charged residues that might stabilize the transition state formed during catalysis. In addition, both enzymes contain a catalytically required metal ion, which is held in position by two aspartates. These similarities do not extend beyond the immediate vicinity of the catalytic centre, with structural divergence seen even in the contacts that stabilize the adenine rings (Fig. 5a, c).

The mechanism through which EF and mAC generate the attacking 3' O[−] ion is also different (Fig. 5a, c). In EF, His351 seems to act as a catalytic base, removing the proton from the 3' hydroxyl. But mACs and their homologues do not have a conserved histidine in their catalytic site and are thought to use a second metal ion for this purpose²¹. mACs share structural homology with the catalytic 'palm domains' of the oligonucleotide polymerases—an enzyme family that functions by two-metal-ion catalysis²¹. Supporting the assertion that mACs work like polymerases, a second metal ion (ion A) adjacent to the first (ion B) has been observed in the structure of Zn²⁺-complexed mAC²¹. Substitution of the metal at site A with a histidine for deprotonation is not unprecedented: two homing endonucleases, *I-PpoI* and *Serratia* nuclease, use this mechanism to catalyse destruction, rather than formation, of a phosphodiester bond^{25,26}.

The ribose in the structure of mAC–ATP- α S is oriented similarly to that in EF–CaM, and the 3' hydroxyl is too far away for effective deprotonation by metal ion A (5.3 Å; ref. 21 and Fig. 5c). Thus, an alternative mechanism for mAC catalysis can be proposed. In the structure of mAC a well ordered water molecule, positioned 3 Å from the 3' hydroxyl of ATP- α S, forms a hydrogen bond with the hydroxyl group of Ser 1028 (Fig. 5c). This water molecule is also 3 Å away from the amino group of Arg 1029—a conserved residue that is crucial for catalysis²⁷. If this 'water' were a hydroxide anion stabilized by the nearby arginine, it could act as a base to draw the proton from the 3' hydroxyl. Alternatively Arg 1029, with its amino group positioned 3.5 Å from the 3' hydroxyl, might either stabilize the 3' O[−] ion or act directly as a catalytic base.

Table 1 Calmodulin activation of wild type and mutant forms of EF

	$V_{\max}$ (s ^{−1})	EC_{50} (nM)†	K_m (mM)‡	$V_{\max}/K_m$
EF*	1,197	1.3	0.21	5,700
Lys346Arg	0.005	32.0	0.55	0.009
Lys353Arg	81	0.8	0.47	173
Lys353Ala	2	1.0	1.38	2
Leu523Ala	559	1.3	0.21	2,657
Lys525Ala	249	256.0	0.26	958
Gln526Ala	698	1.9	0.25	2,792
Val529Ala	436	1.8	0.28	1,557
His577Asn	7	6.1	0.28	23
His577Asp	0.03	8.0	0.31	0.08
Asn583Ala	8	3.3	0.52	16
Asn583Gln	2	1.6	0.72	3
Asn583His	6	1.5	0.86	7
Glu588Ala	50	1.1	0.63	79
Asp590Ala	188	1.1	0.79	238
Asn639Ala	95	2.0	0.30	317
Asp647Ala	4	0.7	0.41	10

*The adenyl cyclase assay was performed with wild-type EF and EF mutants Lys353Arg, Glu588Ala, Asp590Ala, Leu523Ala, Lys525Ala, Gln526Ala, Val529Ala (5 ng), Lys353Ala, His577Asn, His577Asp, Asn583Ala, Asn583Gln, Asn583His, Asn639Ala, Asp647Ala (50 ng), and His577Asn and Lys346Arg (15 μ g) at 30 °C for 10 min in the presence of 5 mM ATP, 10 mM MgCl₂ and 0.05 mM CaCl₂.

† EC_{50} was determined using CaM concentration ranging from 0.01 to 1,000 nM in 3.1- or 10-fold increments and the results were fit to a simple rectangular hyperbola.

‡ K_m was determined using ATP concentrations ranging from 0.06 to 2 mM in 2-fold increments and the results were fit in the same way. Data are representative of at least two experiments.

CaM binding to EF

Calmodulin adopts an extended conformation when bound to EF. The distance between the two CaM lobes is similar to that observed in the crystal structure of extended, Ca²⁺-bound CaM¹². However, residues 79–81 of the central helix are unwound, and the amino-terminal CaM lobe is rotated by 80° with respect to its position in the Ca²⁺-bound crystal structure (Fig. 1). Unwinding of the central helix of CaM is not surprising, as this region is flexible in solution and is unfolded in other peptide–CaM complexes²⁸. In the EF–CaM structure, the C-terminal domain of CaM binds two Ca²⁺ ions and has a similar conformation to that in other Ca²⁺-loaded CaM structures (a C α r.m.s. deviation of 0.62 Å over 61 amino acids)¹². Although 2 mM Ca²⁺ was present in the crystallization drop, the N-terminal lobe of CaM does not contain Ca²⁺ and adopts a closed conformation (a C α r.m.s. deviation of 1.2 Å over 67 amino acids)²⁹, consistent with the finding that inactivation of Ca²⁺-binding site I or II in CaM has no effect on EF activation⁷.

An intricate network of hydrophobic and hydrophilic contacts seems to be responsible for the strong interaction between EF and CaM (dissociation constant K_d = 20 nM)⁷. Over 63 predominantly hydrophobic and basic EF residues contact more than 53 residues from CaM (Fig. 2). These contacts are divided among two segments of the EF sequence: residues 501–540, which are part of domain C_A; and residues 616–798, which correspond to switch C and the helical domain. The first region was identified previously as a CaM-binding site both in EF and in CyaA^{30,31}. This first segment interacts with CaM helices IV, V and VI, and, as discussed below, part of this segment is structurally similar to the CaM interaction sites of other proteins (Fig. 6a). The latter segment represents three CaM-binding surfaces (residues 615–634, 647–672 and 695–721), and the interaction of this part of EF with CaM is consistent with crosslinking data³². These residues contact three of the four EF-hand motifs in CaM (Figs 2 and 6). In particular, EF helices L and M contact the first EF hand of CaM and seem to stabilize the Ca²⁺-free conformation in the N-terminal CaM lobe.

Activation of EF by CaM

In the CaM-free EF structure, C_A, the helical domain and switch C form a contiguous, positively charged surface that is likely to make initial electrostatic contacts with the highly negatively charged CaM (pI = 3.9). The unusually large binding surface between EF and CaM is likely to stabilize the various structural rearrangements that activate EF. Given the extensive interaction observed between the helical domain and the remainder of the protein in the EF-alone structure, it is unlikely that the whole CaM-binding surface becomes exposed as a result of natural structural fluidity. We therefore propose that EF activation by CaM is a multistep process, which is initially mediated by CaM-binding residues that are exposed in the CaM-free state. We identified one such residue, Lys 525, as a binding 'hot spot' through alanine-scanning mutagenesis of EF residues located in the interface between helix H and CaM (Fig. 6 and Table 1). Whereas mutations in nearby residues Lys 523, Gln 526 and Val 529 have little effect on CaM activation, the Lys525Ala mutation markedly increases (200-fold) the CaM concentration for half-maximum response (EC_{50}), with only a small reduction in the velocity of the enzyme-catalysed reaction at infinite concentration of substrate ($V_{\max}$).

In contrast, residues in switch C are likely to bind CaM later in the activation process. Asp 647, for example, is buried in the interdomain interface in the CaM-free state, making a network of hydrogen bonds with atoms from Ser 669, Pro 648 and Ile 649. On CaM binding and rearrangement of switch C, Asp 647 forms a salt bridge with Arg 90 of CaM. The Asp647Ala mutant has a significantly lower $V_{\max}$ but its EC_{50} changes little, suggesting that this residue is more important for enzyme activation than for CaM binding (Fig. 6 and Table 1). Thus, we propose that CaM binds residues from switch A before those of switch C.

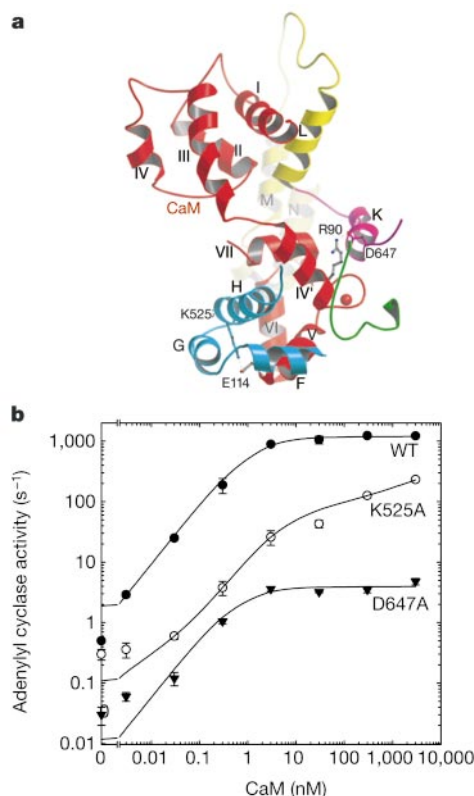


Figure 6 Calmodulin binding of EF. **a**, Ribbon representation of CaM and EF. CaM is in red, and switch A, switch C, the helical domain and C_A/C_B outside the switch regions in EF are in cyan, magenta, yellow and green, respectively. **b**, Mutational analysis of EF residues that contact CaM. The adenylyl cyclase assay was performed with EF (5 ng), Lys525Ala (5 ng) and Asp647Ala (50 ng) at 30 °C for 10 min in the presence of 5 mM ATP, 10 mM MgCl₂ and 0.05 mM CaCl₂.

Switch B (residues 579–591) contains several residues that either bind ATP directly or stabilize the catalytic residues (Fig. 3). The main-chain carbonyl of Thr 579 forms a hydrogen bond with N6 of the adenine base, and Asn 583 forms a crucial interaction with the ribose. Mutating Asn 583 to alanine causes a 180-fold reduction in catalytic activity (Fig. 4c and Table 1). Substitutions of Asn 583 to other hydrogen-bond-forming residues such as glutamine and histidine result in a greater than 400-fold reduction, highlighting the importance of the precise stereochemistry of the Asn 583 side chain (Fig. 4c and Table 1). ATP is locked into the catalytic site by a salt bridge formed between the carboxyl group of Glu 588 and the amino group of Lys 353. Mutations of either residue result in a significant reduction in the catalytic activity (Fig. 4c and Table 1). In addition, Asp 590 forms a salt bridge with Arg 329, a residue that participates directly in catalysis. Mutation of Asp 590 results in more than a 20-fold reduction in the value of $V_{\max}$ /Michaelis constant (K_m) (Fig. 4c and Table 1).

In the EF-alone structure switch B is disordered, but it becomes highly ordered in the EF–CaM complex (Fig. 4). This transition is most probably caused by extensive contacts between switch B and the N-terminal portion of switch C. Notably, Asn 639 forms a hydrogen bond with the main-chain carbonyl of Pro 587. Mutation of Asn 639 to alanine results in a 12-fold reduction in the rate of catalysis (Fig. 4c and Table 1). Thus, we propose that the ordering of switch B occurs as a consequence of the CaM-induced movement of switch C.

CaM binding of other effectors

Oedema factor is homologous to the N-terminal portion of CyaA, the 1,705-residue exotoxin secreted by *B. pertussis*². Despite an

overall sequence identity of only 24%, only 2 of the 15 residues that contact the substrate differ between the two proteins, and these two residues use their main-chain atoms, rather than their side-chain atoms, to contact the substrate. Both EF and CyaA induce an extended conformation of CaM. This conclusion is based on fluorescence resonance energy transfer (FRET) experiments using CaM labelled with a fluorescence donor on one lobe and a non-fluorescent acceptor on the other. Both exotoxins cause a decrease in resonance energy transfer, indicating an increase in the distance between the CaM domains (see Supplementary Information and ref. 7). The K_d of CyaA (residues 1–412) for CaM, obtained from fluorescence titration, is 10 nM. These observations suggest that EF and CyaA use common catalytic and regulatory mechanisms.

The crystal structures of CaM in complex with fragments from myosin light chain kinase (MLCK)¹⁵, CaM kinase II- α (CaMKII α)¹⁶, CaM kinase kinase¹⁷ and the Ca²⁺-sensitive potassium channel¹⁸ define the canonical view of effector–CaM interaction, in which the hydrophobic faces of the two CaM lobes act as a clamp, surrounding a target helix or helices (Fig. 7b, c). The structure of the EF–CaM complex differs markedly from this view. In the EF–CaM complex, the CaM lobes do not act as a clamp. The overall conformation of CaM is extended rather than compact, and the hydrophobic binding surface of the Ca²⁺-free N-terminal lobe is buried within CaM rather than oriented towards the effector target. Only the interaction between the C-terminal lobe of CaM and EF helix H, which makes up 22% of the total CaM-binding surface, resembles the interaction observed with the recognition helix peptides from MLCK and CaMKII (refs 15, 16; and Fig. 7d).

Because peptides derived from several CaM-binding proteins have been used previously to probe the structure of CaM in complex with various CaM effectors^{15–17}, we tested the effect of the EF helix H peptide (residues 521–539) in isolation of the surrounding EF enzyme on the conformation of CaM using FRET. The helix H peptide causes a decrease in interdomain separation in CaM, an effect similar to that induced by an MLCK peptide but opposite to that caused by the whole EF catalytic domain (Supplementary Information). The apparent K_d of the helix H peptide obtained from the titration data was 8.8 μ M, indicating that helix H has roughly 440 times weaker binding compared with EF.

Over 50 CaM effector molecules have been catalogued on the basis of the presence of positively charged amino acids and characteristic 1-5-10 or 1-8-14 spacings of hydrophobic residues in putative CaM-binding helices³³. EF was identified as having a 1-8-14 motif, but in our crystal structure the amino acids with this spacing do not interact with CaM. EF helix H contains a 1-5-10 spacing of hydrophobic residues (Val 529, Lys 533, Ile 538), but these three residues are out of register when superimposed structurally with the 1-5-10 motif observed in the crystal structure of CaM in complex with a CaMKII peptide. Taken together with our FRET data, these observations underscore the difficulty in predicting CaM conformation and CaM interaction with its targets on the basis of sequence homology and peptide studies.

Catalytic activation of other CaM effectors

Removal of an autoinhibitory domain, through conformational changes induced by CaM, is a prevalent mechanism by which CaM activates its targets, including a family of kinases, calcineurin, nitric oxide synthase and phosphodiesterase⁶. Truncation mutants of these proteins lacking an inhibitory domain are constitutively active^{6,34}. Although CaM binding displaces the EF helical domain from the catalytic core, the EF helical domain does not act as an inhibitory domain in the classical sense. A truncation mutant of EF lacking the helical domain is not constitutively active and is significantly less sensitive to CaM activation than wild type⁷. Thus, EF seems to use a different CaM-dependent regulatory mechanism.

There are several mechanisms by which activators are known

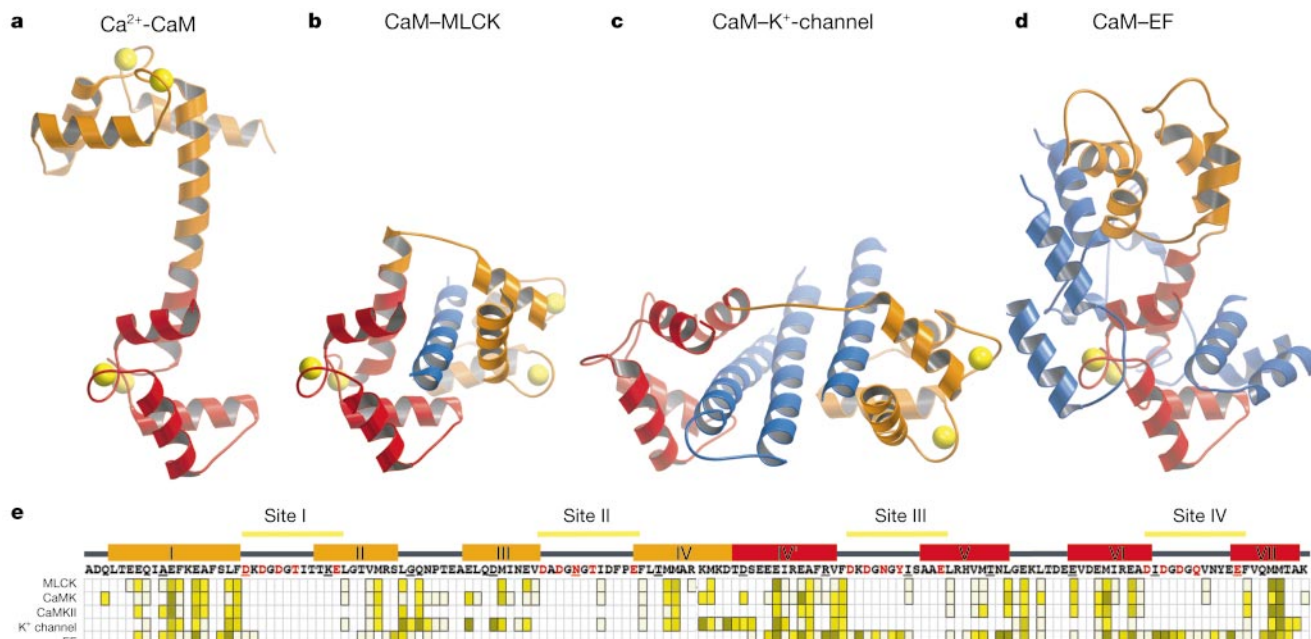


Figure 7 Comparison of CaM conformations and of CaM/effector contacts. **a–d**, Ribbon representations of CaM alone (**a**), CaM in complex with the effector-derived peptides from MLCK (**b**), Ca²⁺-activated K⁺ channel (**c**) and EF (**d**). The C-terminal domain (red) of CaM is aligned on the basis of helix VI of CaM. The CaM central helix and N-terminal domains are in orange, and the effector peptides are in blue. **e**, Sequence and secondary structure of CaM with Ca²⁺-binding sites indicated and Ca²⁺-binding residues shown in

red. Boxes beneath the CaM sequence indicate the effector contact area of each residue in the various structures, using the same colouring scheme as in Fig. 2. The Protein Data Bank codes for the structures are 1CLL for CaM alone, and 1CDL, 1CKK, 1CDM and 1G4Y for CaM in complex with MLCK, CaM kinase, CaMKII, and Ca²⁺-activated K⁺ channel, respectively. Helix designations above the CaM sequence are based on the 1CLL CaM structure.

to activate enzymes. The activation of EF by CaM is conceptually most similar to the activation of G α proteins by regulator of G-protein signalling (RGS)³⁵. In this case, as in EF, the catalytic machinery is present near the site of catalysis, but is disordered in the inactive state. The most surprising aspect of EF activation by CaM is that the conformational switches associated with activation do not involve residues that participate in chemical catalysis. The catalytic base, the metal-binding site, and all of the amino acids thought to be involved in stabilizing the pentavalent transition state are virtually identical in the active and inactive states of the enzyme. These residues are significantly more exposed to solvent (and substrate) in the inactive state than in the active state. The switch that gives rise to a 1,000-fold increase in enzyme activity involves residues that bind and position the substrate, not those that perform the catalytic reaction.

Conclusions

Given the historical importance of prokaryotic toxins to our current understanding of eukaryotic biochemistry, it is fitting that an anthrax protein should help shape our thinking about CaM—one of the most ubiquitous and conserved proteins in vertebrate biology. The structure of EF–CaM is the first example where a functional effector–CaM entity has been characterized structurally, and the complex has the largest protein–protein interface between independently stable molecules observed so far³⁶. By using a combination of structural and biochemical analyses, we have determined details of the complex process whereby CaM, acting at a distance, stabilizes the conformation of the substrate-binding pocket of EF. Although the CaM interactions that we observe are both different and more extensive than those observed previously, members of the EF-hand family to which CaM belongs can interact with many other domains^{37,38} and several CaM effector proteins contain CaM-binding sites that are apparently outside the canonical effector–CaM interface region^{39–43}. Thus, future structural studies of effector–CaM complexes are likely to unveil a rich repertoire of

CaM–effector interactions and activation mechanisms.

EF-like adenyl cyclase exotoxins are found in three pathogens that pose significant human health threats. The number of *B. pertussis* infections has risen recently owing to relaxed immunization; *P. aeruginosa* is a predominant mortality risk for hospitalized people; and *B. anthracis* presents a challenge to non-industrialized countries and has been developed as a biological weapon. Our results indicate that the catalytic sites and the activation mechanism of these three bacterial enzymes are highly conserved but significantly different from those of mAC. Thus, the EF–CaM structure will provide a good starting point for future efforts towards designing structure-based inhibitors of these human pathogens. □

Methods

Structure determination of EF and EF–CaM complex

We purified EF proteins with hexahistidine tags and human CaM as described⁴⁴. Crystals of nucleotide-free EF–CaM complex with and without non-cyclizable ATP in the I222 space group were grown and cryoprotected as described⁴⁴. Crystals of EF alone were grown by the hanging drop method. One microlitre of H₆-EF (40 mg ml^{−1}) in 20 mM Tris-HCl and 50 mM NaCl (pH 7.5) was mixed with 1 μ l of well solution (100 mM citric acid (pH 5.65), 15% glycerol, PEG 550MME at 11.5% (w/v), 220 mM NiSO₄ and 2 mM MgCl₂). In 1–3 weeks, crystals of space group P2₁2₁2 were formed. The crystals were equilibrated for 5 min in cryoprotectant containing well solution plus 30% glycerol. Crystals were frozen in liquid propane and mounted in a nitrogen stream at 100 K. We collected data at APS Biocars 14-BM-C (EF–CaM–3' dATP and EF alone) and NSLS X25 (EF–CaM complex), and processed them with the program HKL or Mosflm (see Supplementary Information)⁴⁵.

The structure of the EF–CaM complex was solved initially by multiwavelength anomalous diffraction (MAD) at 3.8 Å resolution using data collected at APS SBC, and signals from the 27 CaM selenomethionines and 3 Yb³⁺ ions in the asymmetric unit⁴⁴. We improved the initial phase set by including 4.0 Å uranyl fluoride and triethyl lead acetate derivative data⁴⁴. The 3.8 Å experimental phases were extended to 2.75 Å resolution using data from the best crystal of EF–CaM–3' dATP, and iterative cycles of three-fold non-crystallographic symmetry averaging and phase combination in DM⁴⁶. Initial phases for the structures of EF alone (2.6 Å resolution) and EF–CaM (2.95 Å) were obtained by molecular replacement using the program EPMR⁴⁷. We modelled the residues and refined the structure using the programs TURBO-FRODO, O and CNS (Supplementary Information)^{48,49}. The R-factor/*R*_{free} values (%) for the structures of EF–CaM–3' dATP, EF alone and EF–CaM are 21.5/28.6, 22.8/27.6 and 27.8/31.5, respectively. We calculated the

surface areas by GRASP⁵⁰. Figures were generated using Molscript, Povray, and r3dtopov and loosengrass^{50,51}.

Construction and analysis of EF mutants

We used pProExH6-EF as a template to construct plasmids for the expression of mutant proteins, and introduced the desired mutation using the Stratagene Quik-Change kit. The resulting mutation was confirmed by DNA sequencing using the Perkin Elmer Big-Dye kit. We transformed plasmids carrying each mutation into RNaseE-deficient *Escherichia coli* BL21 Star (DE3) containing pUBS520, and purified the mutant proteins as described⁷. *In vitro* adenyl cyclase assays of the purified proteins were done using [α -³²P]ATP as described⁷.

Received 21 September; accepted 23 November 2001.

1. Leppla, S. H. Anthrax toxin edema factor: a bacterial adenylate cyclase that increases cyclic AMP concentrations of eukaryotic cells. *Proc. Natl Acad. Sci. USA* **79**, 3162–3166 (1982).
2. Ladant, D. & Ullmann, A. *Bordetella pertussis* adenylate cyclase: a toxin with multiple talents. *Trends Microbiol.* **7**, 172–176 (1999).
3. Yahr, T. L., Vallis, A. J., Hancock, M. K., Barbieri, J. T. & Frank, D. W. ExoY, an adenylate cyclase secreted by the *Pseudomonas aeruginosa* type III system. *Proc. Natl Acad. Sci. USA* **95**, 13899–13904 (1998).
4. Duesbery, N. S. *et al.* Proteolytic inactivation of MAP-kinase-kinase by anthrax lethal factor. *Science* **280**, 734–737 (1998).
5. Dixon, T. C., Meselson, M., Guillemin, J. & Hanna, P. C. Anthrax. *N. Engl. J. Med.* **341**, 815–826 (1999).
6. Eldik, L. V. & Watterson, D. M. *Calmodulin and Signal Transduction* (Academic, New York, 1998).
7. Drum, C. L. *et al.* An extended conformation of calmodulin induces interactions between the structural domains of adenyl cyclase from *Bacillus anthracis* to promote catalysis. *J. Biol. Chem.* **275**, 36334–36340 (2000).
8. Tang, W. J., Krupinski, J. & Gilman, A. G. Expression and characterization of calmodulin-activated (type I) adenyl cyclase. *J. Biol. Chem.* **266**, 8595–8603 (1991).
9. Deisseroth, K., Heist, E. K. & Tsien, R. W. Translocation of calmodulin to the nucleus supports CREB phosphorylation in hippocampal neurons. *Nature* **392**, 198–202 (1998).
10. Peters, C. & Mayer, A. Ca²⁺/calmodulin signals the completion of docking and triggers a late step of vacuole fusion. *Nature* **396**, 575–580 (1998).
11. DeMaria, C. D., Soong, T. W., Alseikhan, B. A., Alvania, R. S. & Yue, D. T. Calmodulin bifurcates the local Ca²⁺ signal that modulates P/Q-type Ca²⁺ channels. *Nature* **411**, 484–489 (2001).
12. Babu, Y. S. *et al.* Three-dimensional structure of calmodulin. *Nature* **315**, 37–40 (1985).
13. Kuboniwa, H. *et al.* Solution structure of calcium-free calmodulin. *Nature Struct. Biol.* **2**, 768–776 (1995).
14. Finn, B. E. *et al.* Calcium-induced structural changes and domain autonomy in calmodulin. *Nature Struct. Biol.* **2**, 777–83 (1995).
15. Meador, W. E., Means, A. R. & Quiocho, F. A. Target enzyme recognition by calmodulin: 2.4 Å structure of a calmodulin–peptide complex. *Science* **257**, 1251–1255 (1992).
16. Meador, W. E., Means, A. R. & Quiocho, F. A. Modulation of calmodulin plasticity in molecular recognition on the basis of X-ray structures. *Science* **262**, 1718–1721 (1993).
17. Osawa, M. *et al.* A novel target recognition revealed by calmodulin in complex with Ca²⁺/calmodulin-dependent kinase kinase. *Nature Struct. Biol.* **6**, 819–824 (1999).
18. Schumacher, M. A., Rivard, A. F., Bachinger, H. P. & Adelman, J. P. Structure of the gating domain of a Ca²⁺-activated K⁺ channel complexed with Ca²⁺/calmodulin. *Nature* **410**, 1120–1124 (2001).
19. Glaser, P. *et al.* Functional consequences of single amino acid substitutions in calmodulin-activated adenylate cyclase of *Bordetella pertussis*. *EMBO J.* **10**, 1683–1688 (1991).
20. Munier, H. *et al.* The role of histidine 63 in the catalytic mechanism of *Bordetella pertussis* adenylate cyclase. *J. Biol. Chem.* **267**, 9816–9820 (1992).
21. Tesmer, J. J. *et al.* Two-metal-ion catalysis in adenyl cyclase. *Science* **285**, 756–760 (1999).
22. Tang, W.-J. & Hurley, J. H. Catalytic mechanism and regulation of mammalian adenyl cyclases. *Mol. Pharmacol.* **54**, 231–240 (1998).
23. Tesmer, J. J., Sunahara, R. K., Gilman, A. G. & Sprang, S. R. Crystal structure of the catalytic domains of adenyl cyclase in a complex with G_{sa}-GTPγS. *Science* **278**, 1907–1916 (1997).
24. Bieger, B. & Essen, L. O. Structural analysis of adenylate cyclases from *Trypanosoma brucei* in their monomeric state. *EMBO J.* **20**, 433–445 (2001).
25. Galbur, E. A. *et al.* A novel endonuclease mechanism directly visualized for I-PpoI. *Nature Struct. Biol.* **6**, 1096–1099 (1999).
26. Miller, M. D., Cai, J. & Krause, K. L. The active site of *Serratia* endonuclease contains a conserved magnesium–water cluster. *J. Mol. Biol.* **288**, 975–987 (1999).
27. Yan, S. Z., Huang, Z. H., Shaw, R. S. & Tang, W. J. The conserved asparagine and arginine are essential for catalysis of mammalian adenyl cyclase. *J. Biol. Chem.* **272**, 12342–12349 (1997).
28. Lewit-Bentley, A. & Rety, S. EF-hand calcium-binding proteins. *Curr. Opin. Struct. Biol.* **10**, 637–643 (2000).
29. Zhang, M., Tanaka, T. & Ikura, M. Calcium-induced conformational transition revealed by the solution structure of apo calmodulin. *Nature Struct. Biol.* **2**, 758–767 (1995).
30. Munier, H. *et al.* Structural characterization by nuclear magnetic resonance spectroscopy of a genetically engineered high-affinity calmodulin-binding peptide derived from *Bordetella pertussis* adenylate cyclase. *Arch. Biochem. Biophys.* **320**, 224–235 (1995).
31. Glaser, P. *et al.* Identification of residues essential for catalysis and binding of calmodulin in *Bordetella pertussis* adenylate cyclase by site-directed mutagenesis. *EMBO J.* **8**, 967–972 (1989).
32. Labruyere, E. *et al.* Characterization of ATP and calmodulin-binding properties of a truncated form of *Bacillus anthracis* adenylate cyclase. *Biochemistry* **29**, 4922–4928 (1990).
33. Rhoads, A. R. & Friedberg, F. Sequence motifs for calmodulin recognition. *FASEB J.* **11**, 331–340 (1997).
34. Montgomery, H. J., Romanov, V. & Guillemette, J. G. Removal of a putative inhibitory element reduces the calcium-dependent calmodulin activation of neuronal nitric-oxide synthase. *J. Biol. Chem.* **275**, 5052–5058 (2000).
35. Tesmer, J. J., Berman, D. M., Gilman, A. G. & Sprang, S. R. Structure of RGS4 bound to AlF₄-activated G_{qα1}: stabilization of the transition state for GTP hydrolysis. *Cell* **89**, 251–261 (1997).
36. Conte, L. L., Chothia, C. & Janin, J. The atomic structure of protein–protein recognition sites. *J. Mol. Biol.* **285**, 2177–2198 (1999).
37. Meng, W., Sawasdikosol, S., Burakoff, S. J. & Eck, M. J. Structure of the amino-terminal domain of Cbl complexed to its binding site on ZAP-70 kinase. *Nature* **398**, 84–90 (1999).
38. Huang, X. *et al.* Structure of a WW domain containing fragment of dystrophin in complex with β-dystroglycan. *Nature Struct. Biol.* **7**, 634–638 (2000).
39. Krueger, J. K. *et al.* Activation of myosin light chain kinase requires translocation of bound calmodulin. *J. Biol. Chem.* **276**, 4535–4538 (2001).
40. Trehwella, J., Blumenthal, D. K., Rokop, S. E. & Seeger, P. A. Small-angle scattering studies show distinct conformations of calmodulin in its complexes with two peptides based on the regulatory domain of the catalytic subunit of phosphorylase kinase. *Biochemistry* **29**, 9316–9324 (1990).
41. Farrar, Y. J., Lukas, T. J., Craig, T. A., Watterson, D. M. & Carlson, G. M. Features of calmodulin that are important in the activation of the catalytic subunit of phosphorylase kinase. *J. Biol. Chem.* **268**, 4120–4125 (1993).
42. Hayashi, N., Izumi, Y., Titani, K. & Matsushima, N. The binding of myristoylated N-terminal nonapeptide from neuro-specific protein CAP-23/NAP-22 to calmodulin does not induce the globular structure observed for the calmodulin–nonmyristylated peptide complex. *Protein Sci.* **9**, 1905–1913 (2000).
43. Wang, E., Zhuang, S., Kordowska, J., Grabarek, Z. & Wang, C. L. Calmodulin binds to caldesmon in an antiparallel manner. *Biochemistry* **36**, 15026–15034 (1997).
44. Drum, C. L., Shen, Y., Rice, P. A., Bohm, A. & Tang, W. J. Crystallization and preliminary X-ray study of the edema factor exotoxin adenyl cyclase domain from *Bacillus anthracis* in the presence of its activator, calmodulin. *Acta Crystallogr. D* **57**, 1881–1884 (2001).
45. Otwinowski, Z. & Minor, W. Processing of X-ray diffraction data collected in oscillation mode. *Methods Enzymol.* **276**, 307–326 (1997).
46. Cowtan, K. in *Joint CCP4 and ESF-EACBM Newsletter on Protein Crystallography* No. 31, 34–38 (1994).
47. Kissinger, C. R., Gehlhaar, D. K., Smith, B. A. & Bouzida, D. Molecular replacement by evolutionary search. *Acta Crystallogr. D* **57**, 1474–1479 (2001).
48. Jones, T. A., Zou, J. Y., Cowan, S. W. & Kjeldgaard, M. Improved methods for binding protein models in electron density maps and the location of errors in these models. *Acta Crystallogr. A* **47**, 110–119 (1991).
49. Brunger, A. T. *et al.* Crystallography & NMR system: A new software suite for macromolecular structure determination. *Acta Crystallogr.* **54**, 905–921 (1998).
50. Nicholls, A., Sharp, K. A. & Honig, B. Protein folding and association: insights from the interfacial and thermodynamic properties of hydrocarbons. *Proteins Struct. Funct. Genet.* **11**, 281–296 (1991).
51. Esnouf, R. M. Further additions to MolScript version 1.4, including reading and contouring of electron-density maps. *Acta Crystallogr. D* **55**, 938–940 (1999).

Supplementary Information accompanies the paper on *Nature's* website (<http://www.nature.com>).

Acknowledgements

We thank Y. Mabuchi and A. Higbie for technical assistance; X. Yang for computer maintenance; B. Bernat for help with mass spectrometry; P. Gardner for DNA sequencing; J. Geiger, C. Harrison, G. Meinke, K. Moffat, J. Piccirilli, P. Rice, M. Rosner and C.-R. Wang for discussions; and the beamline personnel at NSLS beamline X25 and APS, especially those at BioCars. This work was supported by grants from the NIH NIGMS and American Heart Association to W.-J.T., by an NIH NIDA fellowship to C.L.D., by AHA fellowship to S.-Z.Y., and a grant from the NIH NIAMS to Z.G. A.B. also acknowledges partial support from the American Cancer Society.

Competing interests statement

The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to W.-J.T. (e-mail: wtang@midway.uchicago.edu) or A.B. (e-mail: bohm@bbri.org). Coordinates for EF–CaM–3'dATP, EF alone and EF–CaM have been deposited with the Protein Data Bank (accession codes 1K90, 1K8T and 1K93, respectively).

Selective absorption processes as the origin of puzzling spectral line polarization from the Sun

J. Trujillo Bueno^{*§}, E. Landi Degl'Innocenti[†], M. Collados^{*},
L. Merenda^{*‡} & R. Manso Sainz^{†*}

^{*} Instituto de Astrofísica de Canarias, E-38200 La Laguna, Tenerife, Spain

[†] Dipartimento di Astronomia e Scienza dello Spazio, Università di Firenze,
Largo E. Fermi 2, I-50125 Firenze, Italy

[‡] Dipartimento di Fisica, Università di Trento, Via Sommarive 14, I-38050 Povo,
Italy

[§] Consejo Superior de Investigaciones Científicas, E-28006 Madrid, Spain

Magnetic fields play a key role in most astrophysical systems, from the Sun to active galactic nuclei^{1–3}. They can be studied through their effects on atomic energy levels, which produce polarized spectral lines^{4,5}. In particular, anisotropic radiation ‘pumping’ processes^{6,7} (which send electrons to higher atomic levels) induce population imbalances that are modified by weak magnetic fields^{8,9}. Here we report peculiarly polarized light in the He I 10,830-Å multiplet observed in a coronal filament located at the centre of the solar disk. We show that the polarized light arises from selective absorption from the ground level of the triplet system of helium, and that it implies the presence of magnetic fields of the order of a few gauss that are highly inclined with respect to the solar radius vector. This disproves the common belief^{4,10,11} that population imbalances in long-lived atomic levels are insignificant in the presence of inclined fields of the order of a few gauss, and opens up a new diagnostic window for the investigation of solar magnetic fields.

The Zeeman effect and optical pumping are two mechanisms capable of inducing polarization signals in the spectral lines that originate in the outer layers of stellar atmospheres.

The Zeeman effect¹² requires the presence of a magnetic field, which causes the atomic energy levels to split into different magnetic sublevels. This splitting produces local sources and sinks of light polarization because of the ensuing wavelength shifts of transitions between levels. The Zeeman effect is most sensitive in circular polarization, with a magnitude that scales with the ratio between the Zeeman splitting and the width of the spectral line (which is very much larger than the natural width of the atomic levels). This so-called longitudinal Zeeman effect responds to the line-of-sight component of the magnetic field. In contrast, the transverse Zeeman effect responds to the component of the magnetic field perpendicular to the line of sight, but produces linear polarization signals that are normally negligible for magnetic strengths $B < 100$ G.

Anisotropic radiation pumping^{6,7} produces atomic level polarization—that is, population imbalances and quantum interferences between the sublevels of degenerate atomic levels (Fig. 1). The presence of a magnetic field is not necessary for the operation of such optical pumping processes, which can be particularly efficient in creating atomic polarization if the depolarizing rates from elastic collisions are sufficiently low. As clarified below, the mere presence of atomic polarization of the type illustrated in Fig. 1 implies local sources and sinks of linear polarization. The Hanle effect^{8,9} (Fig. 1) modifies the atomic polarization of the unmagnetized reference case, and gives rise to a complicated magnetic field dependence of the linear polarization of the scattered light, which is being increasingly applied as a diagnostic tool for weak magnetic fields in astrophysics.

It is often assumed that the observable effects of atomic polarization of long-lived atomic levels are strongly suppressed by Hanle

depolarization in the presence of solar magnetic fields that are highly inclined with respect to the solar radius and have strengths in the gauss range^{4,10,11}. As some of the many ‘enigmatic’ signals of scattering polarization that have been observed within the edge of the solar disk^{13–15} have been shown to be due to ground state atomic polarization^{9,16,17}, and because milligauss (or weaker) magnetic fields are believed to be very rare in the highly conductive solar plasma, it has been concluded that the field must be oriented fairly close to the radial direction wherever the signatures of lower-level atomic polarization are observed¹¹. Although this conclusion is probably valid for the observations of scattering polarization in the Na I D₁ line¹⁶, it should not be generalized to all cases in which the signatures of lower-level atomic polarization are observed. Moreover, the observable effects of lower-level atomic polarization, whether the Hanle effect destroys or creates linear polarization signals in spectral lines, depend on the scattering geometry. To clarify the situation, it is necessary to investigate carefully to what extent observable effects of the atomic polarization of long-lived

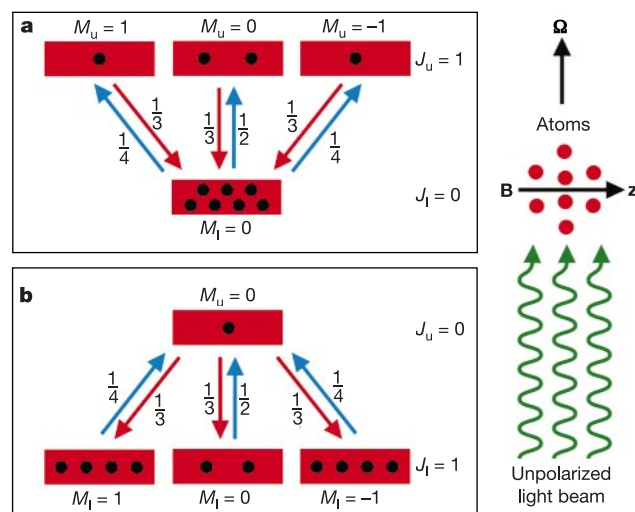


Figure 1 Anisotropic radiation pumping and the Hanle effect. An unpolarized radiation field can induce population imbalances and quantum interferences (or coherences) between the sublevels of degenerate atomic levels (that is, atomic polarization) if the illumination of the atomic system is anisotropic. For example, upper-level population pumping occurs when some upper-state sublevels have more chance of being populated than others (a). On the contrary, lower-level depopulation pumping occurs when some lower-state sublevels absorb light more strongly than others (b). We note that line transitions between levels having other total angular momentum values (for example, $J_l = J_u = 1$ or with $J_l = 1$ and $J_u = 2$) permit the transfer of atomic polarization between both levels via repopulation pumping (for example, lower-level atomic polarization can result simply from the spontaneous decay of a polarized upper level). The Hanle effect is the modification of the atomic polarization of degenerate atomic levels caused by the action of a magnetic field such that its corresponding Zeeman splitting is comparable to the inverse lifetime (or natural width) of the degenerate atomic level under consideration. For the Hanle effect to operate, the magnetic field vector (**B**) has to be significantly inclined with respect to the symmetry axis (**Ω**) of the pumping radiation field. The formula used to estimate the maximum magnetic field intensity B (in G) to which the Hanle effect can be sensitive is $10^6 Bg \approx 1/t_{\text{life}}$, where g and t_{life} are, respectively, the Landé factor and the lifetime (in seconds) of the given atomic level. In a reference frame whose z -axis (that is, the quantization axis of total angular momentum) is parallel to the direction of **B**, the population imbalances turn out to be insensitive to the magnetic field, while the coherences are reduced and dephased as the magnetic strength is increased. This so-called magnetic field reference frame is the one we have chosen here while visualizing the induced population imbalances for the ‘strong field’ case (that is, the case for which the coherences are negligible). We note that the atomic polarization of a given atomic level depends sensitively on the complexity of the assumed atomic model.

atomic levels can not only survive partial Hanle destruction but also even be enhanced by horizontal magnetic fields in the gauss range.

Such investigations can be performed in magnetized astrophysical plasmas, such as those in solar prominences and filaments. These features are relatively cool and dense ribbons of plasma embedded in the 10^6 K solar corona. The ribbons are thought to be confined by the action of highly inclined magnetic fields with strengths in the gauss range¹⁸. Prominences and filaments are in fact the same phenomenon but observed in different circumstances. Both prominences and filaments absorb the photons from the underlying solar photosphere, and re-emit them in all directions. But prominences are observed outside the visible outer edge of the Sun (that is, against the dark background of the sky), while filaments are the same types of structures observed against the bright background of the solar disk. Therefore, we see emission lines in prominences, but absorption lines in filaments.

Figure 2 contrasts prominences and filaments, and illustrates our theoretical prediction concerning the expected linear polarization of a line transition with $J_l = 1$ and $J_u = 0$, where J_l and J_u are the angular momentum quantum numbers of the lower and upper

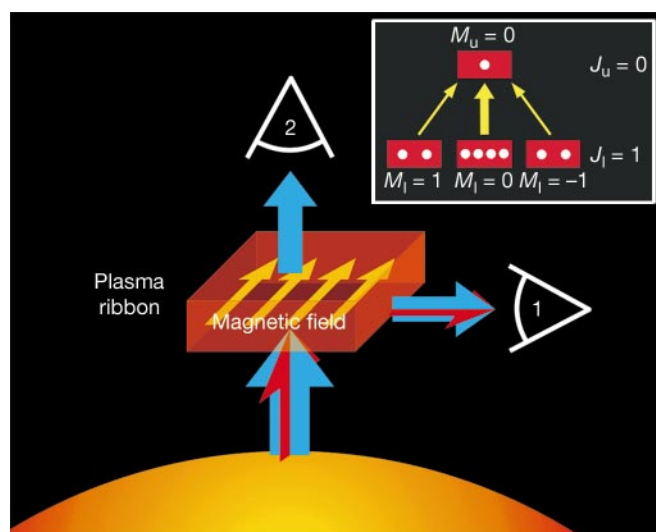


Figure 2 The solar prominence case versus the solar filament case. In a prominence located in the plane of the sky at a given distance above the visible outer edge of the Sun we see the result of 90° scattering events, whereas in a filament situated exactly at the centre of the solar disk we see the result of forward-scattering events. This figure illustrates these two cases by considering observations of a magnetized plasma ribbon using polarimeters at positions 1 (the prominence case) and 2 (the filament case). The figure refers exclusively to the 'blue' line of the He I 10,830-Å multiplet, which is a line transition with $J_l = 1$ and $J_u = 0$. As the upper level cannot carry any atomic polarization, the spontaneously emitted radiation which follows the anisotropic radiative excitation is virtually unpolarized. For this reason, the observer at position 1 sees that the fluorescently scattered beam is unpolarized. However, if the lower level is polarized as indicated in the inset, then the transmitted beam seen by the observer at position 2 will have an excess of linear polarization perpendicular to the direction of the horizontal magnetic field, simply because the $\Delta M = 0$, or π transitions, absorb more efficiently than the $\Delta M = \pm 1$, or σ transitions. This selective absorption mechanism²⁰ is called dichroism because the plasma is behaving as a dichroic medium (that is, the absorption coefficient in the line transition depends on the polarization of the radiation). We note that repopulation pumping is important in polarizing the ground level of the triplet system of He I, as our calculations are based on a realistic multiterm atomic model and not on the two-level model atom considered in Fig. 1. This is why the lower-level polarization shown in the inset is different from Fig. 1b. So for $1 \rightarrow 0 \rightarrow 1$ scattering processes we expect to observe virtually zero linear polarization in optically thin prominences, but a sizeable linear polarization signal in filaments if a significant amount of lower-level atomic polarization is present.

level, respectively. In fact, there are two mechanisms by means of which atomic level polarization can generate linear polarization signals in spectral lines: the first is due to the emission process (that is, to the atomic polarization of the upper level), while the second is subtly related to the absorption process (that is, to the atomic polarization of the lower level). In general, the first mechanism (caused exclusively by the spontaneous emission events that follow the anisotropic radiative excitation) is the only one that is taken into account^{4,10,11,14,19}. The role played by the atomic polarization of the lower level in producing linear polarization via the absorption process²⁰ has never been considered seriously.

We have observed the intensity and polarization of the He I 10,830-Å multiplet in a variety of solar prominences and filaments using the Tenerife Infrared Polarimeter²¹ attached to the Vacuum Tower Telescope²². This multiplet originates between a lower term (2^3S_1) and an upper term ($2^3P_{2,1,0}$). Therefore, it has three spectral lines²³: a 'blue' line at 10,829.09 Å (with $J_l = 1$ and $J_u = 0$), and two 'red' lines at 10,830.25 Å (with $J_u = 1$) and at 10,830.34 Å (with $J_u = 2$) which appear blended at the plasma temperatures of prominences and filaments. As explained in Fig. 2 legend, detection of a significant linear polarization signal in the 'blue' line ($J_l = 1$ and $J_u = 0$) would be due to the atomic polarization of the lower level with $J_l = 1$. We note that this lower level is metastable—that is, it is a relatively long-lived atomic level whose atomic polarization is vulnerable (via the ground level Hanle effect^{24,9}) to magnetic fields of very low intensity ($\sim 10^{-3}$ G).

We first consider prominences. The data points in Fig. 3 show the four Stokes parameters observed in a prominence. As expected from the theoretical prediction of Fig. 2, the 'blue' line of the He I 10,830-Å multiplet does not show any significant linear polarization, which implies that this particular prominence has a very small optical thickness along the line of sight. However, there are very significant Stokes Q and U signals around the wavelengths of the blended 'red' components. These linear polarization signals are the observational signature of the atomic polarization of the upper levels with $J_u = 2$ and $J_u = 1$. Note that there are sizeable circular polarization signals in both the 'blue' and 'red' components. They are the result of the longitudinal Zeeman effect. Their detection is essential for the determination of the intensity of the magnetic field because for fields larger than only a few gauss the He I 10,830-Å multiplet enters into the saturated Hanle effect regime for the upper level, where the linear polarization signals are sensitive only to the orientation of the magnetic field vector.

The solid lines in Fig. 3 show the results of our theoretical modelling, taking into account the influence of ground-level polarization. From the fit to the observation we infer a magnetic field of about 40 G, inclined at 31° to the radial direction through the observed point. The dotted lines in Fig. 3 show what happens if we carry out the calculation with this magnetic field vector, but assuming a completely unpolarized ground level. In the present prominence case, the significant difference with respect to the solid-line calculation is solely the result of the feedback that the existing ground-level polarization is producing on the atomic polarization of the upper levels with $J_u = 2$ and $J_u = 1$. We note that a magnetic field diagnostic of solar prominences that neglects the influence of ground-level polarization would imply a significant error ($\sim 10^\circ$) in the field orientation and in the magnetic strength (of a few gauss).

We now turn our attention to filaments. The data points in Fig. 4 show the observed Stokes parameters in a solar filament that was situated exactly at the centre of the solar disk on 2 June 2001. We selected this coronal filament to demonstrate that linear polarization signals can be produced even at the very centre of the solar disk where we meet the forward scattering case. As seen in the figure, the 'blue' line now shows a very significant linear polarization signal with a negative Stokes Q amplitude, which is of the same order of magnitude as the positive Stokes Q amplitude observed in the 'red' blended component. The observed linear polarization signals are

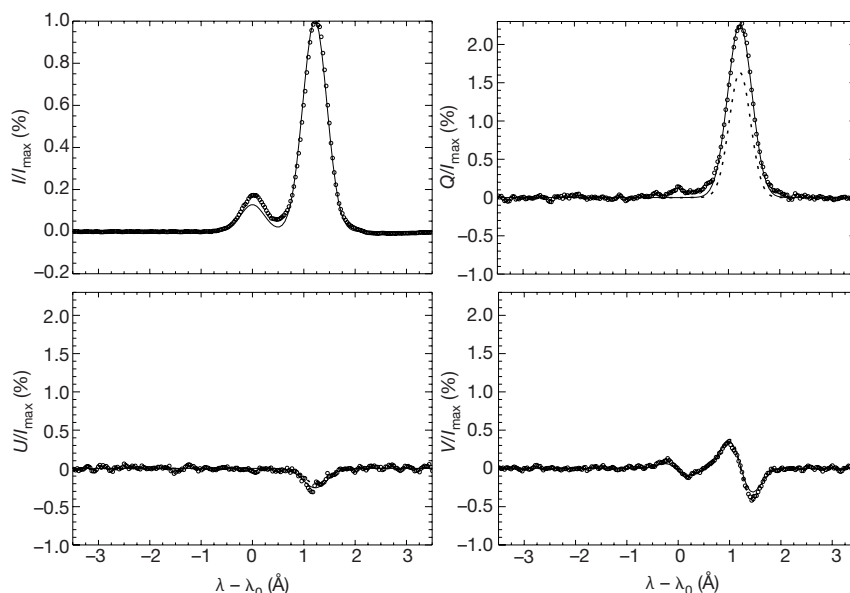


Figure 3 Prominence case: observation versus theory. Spectropolarimetric observation of a solar prominence (circles) versus theoretical modelling taking into account the influence of ground-level atomic polarization (solid line) or neglecting it (dotted line). Our modelling assumes that the He I atoms, lying at a given height (h) above the solar photosphere, are illuminated by an unpolarized and spectrally flat radiation field. It is based on the quantum theory of the generation and transfer of polarized radiation^{25,26}, which we have applied describing the He I atoms in the incomplete Paschen-Back effect regime²⁷. The Stokes I -parameter quantifies the total intensity of the observed light, the Stokes Q and U parameters represent the degree of linear polarization along two reference axes that form an angle of 45° between them, while Stokes V quantifies the degree of circular

polarization²⁸. The observed prominence region had a projected height on the plane of the sky of $20''$ over the visible edge of the solar disk (that is, $h \approx 15,000$ km). The fit to the observations was done assuming a magnetic field vector with intensity $B = 40$ G, inclination $\theta_B = 31^\circ$, and azimuth $\chi_B = 176^\circ$. Note that $\lambda_0 = 10,829.09 \text{ \AA}$ is the line centre wavelength of the 'blue' component of the He I 10,830- $\text{\AA}$ multiplet. The positive reference direction for Stokes Q is perpendicular to the solar radius vector through the observed point. The Stokes profiles are normalized to the maximum line-core intensity of the 'red' emission line. For this particular geometry of scattering, the determination of the magnetic field is ambiguous. The alternative determination $B = 40$ G, $\theta'_B = 180^\circ - \theta_B = 149^\circ$, $\chi'_B = -\chi_B = -176^\circ$, gives the same theoretical curve.

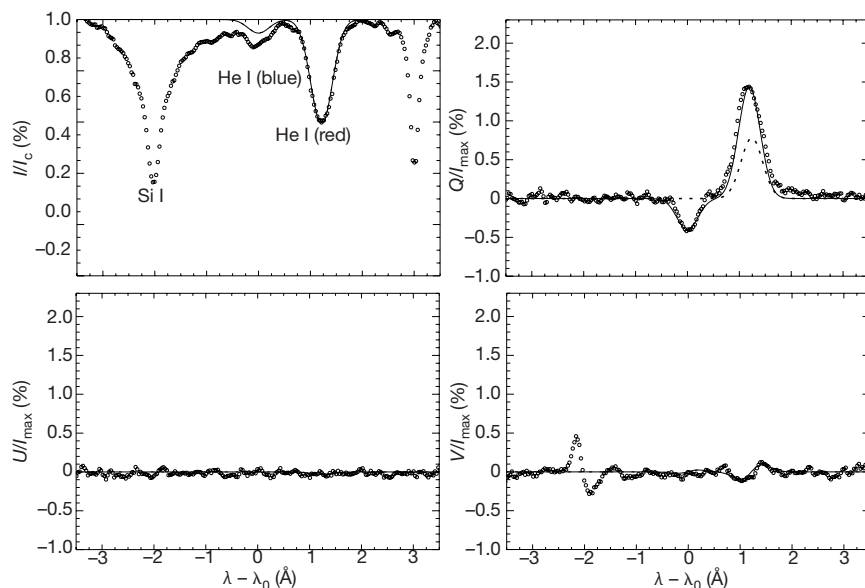


Figure 4 Filament case: observation versus theory. Spectropolarimetric observation of a solar filament located at the disk centre (circles) versus theoretical modelling taking into account the influence of ground-level atomic polarization (solid line) or neglecting it (dotted line). Our choice for the positive reference direction of the observed Stokes Q parameter is the one which minimizes Stokes U . The solid-line fit has been achieved assuming $B = 20$ G, $\theta_B = 105^\circ$ and a height $h = 40''$ (that is, about 30,000 km) above the solar surface. The positive reference direction for the theoretical Stokes Q profile is parallel to the projection of the magnetic field vector on the solar surface. Therefore, negative Stokes Q values indicate that the linear polarization of the observed beam is

perpendicular to the magnetic field of the filament plasma, as illustrated in Fig. 2. The dotted line neglects the influence of ground-level polarization, but takes into account the (negligible) contribution of the transverse Zeeman effect. The Stokes I profile is normalized to the local continuum intensity, while Stokes Q , U and V are normalized to the maximum line-core depression (from the continuum level) of the Stokes I profile of the 'red' absorption line. The weakness of the Stokes V signal, which is caused by the longitudinal Zeeman effect, arises because the magnetic field vector is almost parallel to the solar surface. We note that, for the particular geometry of this observation, an ambiguity of 180° is present in the determination of the azimuth of the magnetic field vector.

entirely due to the Hanle effect operating at disk centre. This can be possible only if there exists a magnetic field with a significant inclination to the radial direction through the observed point, otherwise the polarization at disk centre would be zero for reasons of symmetry. The very existence of a sizeable Stokes Q signal in the 'blue' line demonstrates that the 3S_1 ground level is significantly polarized.

The solid line in Fig. 4 shows the result of our theoretical modelling of the Hanle effect at disk centre taking into account the influence of ground level polarization. From the fit to the observation we infer a magnetic field of 20 G, inclined by about 105° to the radial direction through the observed point and with a horizontal component at an angle of about 10° in the clockwise direction with respect to the axis of the filament. The agreement with the spectropolarimetric observation is notable. It demonstrates that the ground-level Hanle effect is operating in the outer solar atmosphere, producing very significant linear polarization signals by selective absorption from the unevenly populated magnetic sublevels of a long-lived atomic level.

Our results show that the atomic polarization of long-lived levels, which is induced by optical pumping processes, generates observable polarization signatures due to the highly inclined magnetic fields with strengths in the gauss range that are characteristic of solar prominences. Instead of destroying the atomic polarization, the magnetic fields produce (via the Hanle effect) linear polarization signals that are of the same order of magnitude as those caused by the atomic polarization of the short-lived excited states. Moreover, our results provide observational evidence of the operation of an atomic effect that may have diagnostic use in several astrophysical contexts. It concerns the creation of linear polarization signals in spectral lines induced by magnetic fields in forward scattering and by dichroism. These phenomena reveal unfamiliar aspects of the Hanle effect, and open up a new diagnostic window on the investigation of the magnetism of the outer solar atmosphere (chromosphere, transition region and corona). □

Received 22 August; accepted 20 November 2001.

1. Parker, E. N. *Cosmical Magnetic Fields* (Clarendon, Oxford, 1979).
2. Schrijver, C. J. & Zwaan, C. *Solar and Stellar Magnetic Activity* (Cambridge Univ. Press, Cambridge, 2000).
3. Blandford, R. et al. in *Astrophysical Spectropolarimetry* (eds Trujillo Bueno, J., Moreno-Inserit, F. & Sánchez, F.) 177–223 (Cambridge Univ. Press, Cambridge, 2002).
4. Stenflo, J. O. *Solar Magnetic Fields: Polarized Radiation Diagnostics* (Kluwer, Dordrecht, 1994).
5. Landi Degl'Innocenti, E. in *Astrophysical Spectropolarimetry* (eds Trujillo Bueno, J., Moreno-Inserit, F. & Sánchez, F.) 1–53 (Cambridge Univ. Press, Cambridge, 2002).
6. Cohen-Tannoudji, C. & Kastler, A. Optical pumping. *Prog. Opt.* **5**, 3–81 (1966).
7. Happer, W. Optical pumping. *Rev. Mod. Phys.* **44**, 169–249 (1972).
8. Hanle, W. Über magnetische Beeinflussung der Polarisation der Resonanzfluoreszenz. *Z. Phys.* **30**, 93–105 (1924).
9. Trujillo Bueno, J. in *Advanced Solar Polarimetry: Theory, Observations and Instrumentation* (ed. Sigwarth, M.) 161–195 (ASP Conf. Series, Astronomical Society of the Pacific, San Francisco, 2001).
10. Stenflo, J. O. Quantum interferences, hyperfine structure, and Raman scattering on the Sun. *Astron. Astrophys.* **324**, 344–356 (1997).
11. Stenflo, J. O. in *Advanced Solar Polarimetry: Theory, Observations and Instrumentation* (ed. Sigwarth, M.) 97–108 (ASP Conf. Series, Astronomical Society of the Pacific, San Francisco, 2001).
12. Condon, E. U. & Shortley, G. H. *The Theory of Atomic Spectra* (Cambridge Univ. Press, Cambridge, 1979).
13. Stenflo, J. O. & Keller, C. The second solar spectrum: a new window for diagnostics of the Sun. *Astron. Astrophys.* **321**, 927–934 (1997).
14. Lin, H., Penn, M. J. & Kuhn, J. R. He I 10830 line polarimetry: A new tool to probe the filament magnetic fields. *Astrophys. J.* **493**, 978–995 (1998).
15. Stenflo, J. O., Keller, C. & Gandorfer, A. Anomalous polarization effects due to coherent scattering on the sun. *Astron. Astrophys.* **355**, 789–804 (2000).
16. Landi Degl'Innocenti, E. Evidence against turbulent and canopy-like magnetic fields in the solar chromosphere. *Nature* **392**, 256–258 (1998).
17. Trujillo Bueno, J. in *Solar Polarization* (eds Nagendra, K. N. & Stenflo, J. O.) 73–96 (Kluwer Academic, Dordrecht, 1999).
18. van Ballegoijen, A. A. Solar prominence models. In *Encyclopedia of Astronomy and Astrophysics* (ed. Murdin, P.) 2703–2707 (Nature Publishing Group, London, 2001); also at <http://www.ency-astro.com> (2001).
19. Faurobert-Scholl, M. Hanle effect with partial frequency redistribution. I. Numerical methods and first applications. *Astron. Astrophys.* **246**, 469–480 (1991).
20. Trujillo Bueno, J. & Landi Degl'Innocenti, E. Linear polarization due to lower-level depopulation pumping in stellar atmospheres. *Astrophys. J.* **482**, L183–L186 (1997).

21. Martínez Pillet, V. et al. in *High Resolution Solar Physics: Theory, Observations, and Techniques* (eds Rimmele, T. R., Balasubramaniam, K. S. & Radick, R. R.) 264–272 (ASP Conf. Series 183, Astronomical Society of the Pacific, San Francisco, 1999).
22. Soltau, D. in *The Role of Fine-Scale Magnetic Fields on the Structure of the Solar Atmosphere* (eds Schröter, E. H., Vázquez, M. & Wyller, A. A.) 362–366 (Cambridge Univ. Press, Cambridge, 1987).
23. Radzig, A. A. & Smirnov, B. M. *Reference Data on Atoms, Molecules, and Ions* (Springer, Berlin, 1985).
24. Landolfi, M. & Landi Degl'Innocenti, E. Resonance scattering and the diagnostic of very weak magnetic fields in diffuse media. *Astron. Astrophys.* **167**, 200–206 (1986).
25. Landi Degl'Innocenti, E. The determination of vector magnetic fields in prominences from the observations of the Stokes profiles in the D_3 line of helium. *Solar Phys.* **79**, 291–322 (1982).
26. Landi Degl'Innocenti, E. Polarization in spectral lines: I. A unifying theoretical approach. *Solar Phys.* **85**, 3–31 (1983).
27. Bommier, V. Quantum theory of the Hanle effect. *Astron. Astrophys.* **87**, 109–120 (1980).
28. Chandrasekhar, S. *Radiative Transfer* (Dover, New York, 1960).

Acknowledgements

We thank R. Casini and J. O. Stenflo for discussions on quantum electrodynamics and for helping with the presentation of this Letter. The German Vacuum Tower Telescope is operated by the Kiepenheuer Institut at the Observatorio del Teide of the Instituto de Astrofísica de Canarias (IAC). The Tenerife Infrared Polarimeter has been developed by the IAC. We also acknowledge the support of the European Solar Magnetometry Network and of the Spanish Plan Nacional de Astronomía y Astrofísica.

Correspondence and requests for materials should be addressed to J.T.B. (e-mail: jtb@ll.iac.es).

A laboratory analogue of the event horizon using slow light in an atomic medium

Ulf Leonhardt

School of Physics and Astronomy, University of St Andrews, North Haugh, St Andrews, Fife KY16 9SS, UK

Singularities underlie many optical phenomena¹. The rainbow, for example, involves a particular type of singularity—a ray catastrophe—in which light rays become infinitely intense. In practice, the wave nature of light resolves these infinities, producing interference patterns. At the event horizon of a black hole², time stands still and waves oscillate with infinitely small wavelengths. However, the quantum nature of light results in evasion of the catastrophe and the emission of Hawking radiation³. Here I report a theoretical laboratory analogue of an event horizon: a parabolic profile of the group velocity⁷ of light brought to a standstill in an atomic medium^{4–6} can cause a wave singularity similar to that associated with black holes. In turn, the quantum vacuum is forced to create photon pairs with a characteristic spectrum, a phenomenon related to Hawking radiation³. The idea may initiate a theory of 'quantum' catastrophes, extending classical catastrophe theory^{8,9}.

Optical media govern the propagation of light. Media are usually transparent substances such as glass or water, but empty yet curved space is a medium as well¹⁰. Certain material media can be manipulated to give them extraordinary optical properties. Inside such media light may propagate with a negative¹¹ or very low¹² group velocity⁷, or may be brought to a standstill^{4–6}. In a medium with electromagnetically induced transparency¹³ (EIT), an external control beam dictates the group velocity v_g of a second and weaker probe beam, in order to slow down the probe light^{4–6,12}. Once the first beam has gained control, the group velocity of the second beam is essentially proportional to the control intensity I_c , even in the limit when I_c vanishes¹⁴.

Imagine that the control beam illuminates the EIT medium from above, see Fig. 1. Initially, the control intensity is uniform, but then

the control light develops a dark stripe. The stripe should continue down through the medium as an interface Z of zero intensity I_c where, consequently, the group velocity of any potential probe light vanishes. In the following, I show theoretically that the interface Z forms the optical analogue of an event horizon. The formation of the horizon should trigger a quantum catastrophe that results in continuous emission of slow-light quanta out of the vacuum. (Details of the calculations will be published elsewhere.) I will first analyse classical waves of slow light, and then turn to the quantum theory.

Assume that the interface Z of zero control intensity I_c is sufficiently flat that the optical properties generated do not vary much in the spatial directions parallel to Z . Consider a line z orthogonal to Z . Over a small fraction of a characteristic length a , the group-velocity profile of the slow probe light is parabolic

$$v_g \sim c \frac{z^2}{a^2} \quad (1)$$

because I_c increases quadratically in the vicinity of a zero. As usual, c denotes the speed of light in vacuum. For simplicity I concentrate on slow-light waves $\varphi(t, z)$ that propagate in the z direction only, and ignore their polarizations.

To predict the propagation properties of slow light I use the specific physics of EIT illustrated in Fig. 2. Equivalently and more generally, I translate the phenomenological dispersion relation¹⁵ of slow light with frequencies close to the EIT resonance ω_0 into a wave equation that is subject to the principle of least action of the canonical formalism¹⁶:

$$\left(\frac{\partial}{\partial t} (1 + \alpha) \frac{\partial}{\partial t} - c^2 \frac{\partial^2}{\partial z^2} + \alpha \omega_0^2 \right) \varphi = 0 \quad (2)$$

After the formation of the interface Z , the group index α has developed a quadratic singularity:

$$\alpha = c/v_g - 1 = \frac{a^2}{z^2} \quad (3)$$

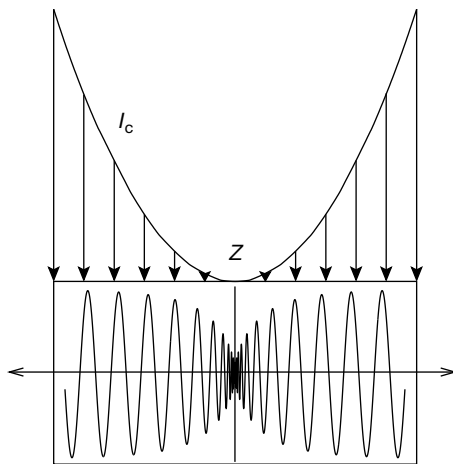


Figure 1 Schematic diagram of the proposed experiment. A beam of control light with intensity I_c generates electromagnetically induced transparency¹³ in a medium, strongly modifying its optical properties for a second field of slow light. When an initially uniform control intensity is turned into the parabolic profile shown in the figure, the slow-light field suffers a quantum catastrophe. To slow-light waves, the interface Z of zero control intensity cuts space into two disconnected regions and creates a logarithmic phase singularity, in analogy to the effect¹⁹ of an event horizon². The quantum vacuum of slow light cannot occupy such catastrophic waves. In turn, pairs of slow-light quanta, propagating in opposite directions away from Z , are emitted with a characteristic spectrum. The waves shown below the intensity profile refer to the emitted light with the modes w_R and w_L of equation (7).

As a consequence, slow light propagates independently on the two sides of the interface Z and can never cross Z , because, in mathematical terms, it is possible to multiply any solution φ with the step function $\Theta(\pm z)$ and still solve the wave equation (2). On either side of Z a slow-light pulse can be decomposed into monochromatic waves—that is, into stationary solutions of the wave equation (2)

$$\varphi = \sqrt{z} J_{\pm \nu}(kz) e^{-i\omega t}, \quad \nu = \sqrt{1/4 - a^2(k^2 - k_0^2)} \quad (4)$$

expressed in terms of the Bessel functions¹⁷ J_ν and the wavenumbers $k = \omega/c$. Two cases emerge. First, when $4a^2(\omega^2 - \omega_0^2) \leq c^2$, the index ν is real. In this case, the incident waves are totally reflected away from the interface Z , as inferred from the behaviour of the Bessel functions¹⁷ for large kz . In the other case, $4a^2(\omega^2 - \omega_0^2) > c^2$, the index is imaginary,

$$\nu = i\mu, \quad \mu = \sqrt{a^2(k^2 - k_0^2) - 1/4} \quad (5)$$

and the reflected and incident waves are not balanced. The remaining transmitted light is trapped at the interface Z , because here¹⁷:

$$\varphi \propto \zeta^{i\mu+1/2} e^{-i\omega t} = \sqrt{\zeta} e^{i\mu \ln \zeta - i\omega t}, \quad \zeta = kz \quad (6)$$

Close to Z , the slow-light intensity reduces with reducing distance z and the phase $\mu \ln(kz)$ becomes infinite, causing the light to oscillate with linearly decreasing wavelength⁷ $(2\pi/\mu)z$. Waves ‘freeze’ near the interface Z . A process that creates an interface where waves separate and develop a logarithmic phase singularity¹⁸ may be termed ‘wave catastrophe’.

Close to the event horizon of a black hole², an outside observer would see a similar behaviour of waves¹⁹. The horizon cuts space into two disconnected parts. All motion freezes near the horizon where time seems to stand still. Waves develop a logarithmic phase singularity. Yet an observer falling inwards could pass the horizon without noticing anything unusual. This characteristic difference in perception has a profound consequence, because the quantum vacuum behaves in a similar fashion to a fluid that shares the fate of the inward-falling observer^{19–21}. Consequently, the quantum vacuum must not occupy the waves seen by the outside observer.

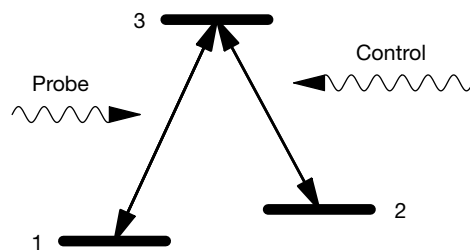


Figure 2 Physical processes behind electromagnetically induced transparency¹³ (EIT). The figure shows the relevant energy levels of each atom constituting the EIT medium. The control light couples two excited states $|2\rangle$ and $|3\rangle$, and thus influences the optical transition between the ground state $|1\rangle$ and the level $|3\rangle$ brought about by the probe light. Initially, the control light, being sufficiently strong, prepares each atom in a pure state $|\psi\rangle$ called a dark state^{13,14}. When the control and probe field strengths vary, the atoms remain in dark states as long as level $|3\rangle$ is not sufficiently populated. Up to a normalization and phase factor, the dark state $|\psi\rangle$ is proportional to $|1\rangle - (\Omega_p/\Omega_c)|2\rangle + 2(N_0/\Omega_c^2)\partial(\Omega_p/\Omega_c)/\partial t|3\rangle$ with $N_0^{-2} = 1 + |\Omega_p/\Omega_c|^2$, described here in an interaction picture with respect to the atomic transition frequencies ω_{32} and $\omega_{31} = \omega_0$. The field strengths of the probe and control light are given in terms of the local Rabi frequencies²² Ω_p and Ω_c . The induced dipole moments of the atoms in dark states generate a matter polarization that influences the propagation of the probe light. When $|\Omega_p|^2$ is much less than $|\Omega_c|^2$ the probe light obeys the linear wave equation (2) with a group index¹⁵ $\alpha = c/v_g - 1$ that is inversely proportional¹⁴ to $|\Omega_c|^2$. The less intense the control field, the lower is the group velocity v_g . When $|\Omega_p|^2$ is comparable with $|\Omega_c|^2$ or larger, nonlinear optical effects occur.

In other words, this observer does not see a vacuum. Instead, the observer detects the quanta of Hawking radiation^{3,19}.

Consider the quantum physics of our wave catastrophe. According to quantum field theory¹⁶, waves are potential particle carriers called modes. Modes describe the spatial-temporal fields of single quanta and, therefore, they are normalized with respect to a characteristic scalar product²⁰. I normalize the waves of equation (4) with the imaginary index of equation (5) to find the set of modes:

$$\begin{aligned} w_R &= \frac{1}{\sqrt{1 - e^{-2\pi\mu}}}(u_R^+ - e^{-\pi\mu}u_R^-), & w_L(z) &= w_R(-z), \\ u_R &= u_R^-, & u_L(z) &= u_R(-z), \\ u_R^\pm &= \frac{\Theta(z)}{\sqrt{2c}}e^{-\mu\pi/2}\sqrt{k_0z}J_{\pm i\mu}(kz)e^{-i\omega t} \end{aligned} \quad (7)$$

The step function Θ indicates that the R/L modes exist either on the right or on the left side of the horizon Z . I have chosen the w modes such that they appear as outgoing plane waves far away from Z . These are the modes that carry detectable quanta.

Complex analysis¹⁸ is an imaginative mathematical method to understand real physics. Regard hypothetically the distance z and the time t as complex variables¹⁸. The modes of equation (7) are non-analytic functions¹⁸ of z , because they vanish¹⁸ on one of the sides of the horizon. However, the quantum vacuum should occupy a set of analytic modes, because the formation of the horizon is a dynamic process¹⁴. Initially, the quantum vacuum occupies analytic modes such as packets of plane waves. The process conserves the analyticity in z , even after the wave catastrophe has occurred. Furthermore, the light waves are analytic on the lower half of the complex t plane on at least one side of the horizon, see Fig. 3. It is possible to prove that at any arbitrary time t_0 after the catastrophe, the vacuum modes are certain combinations of the detector modes

$$\begin{aligned} v_R &= \frac{w_L}{e^{\pi\mu} + e^{-\pi\mu}} - iw_R - ie^{-2i\omega t_0} \frac{w_R^*}{e^{\pi\mu} + e^{-\pi\mu}}, \\ v_R^\dagger &= \frac{1}{\sqrt{1 + e^{2\pi\mu}}}(u_R + ie^{\pi\mu}u_L) \end{aligned} \quad (8)$$

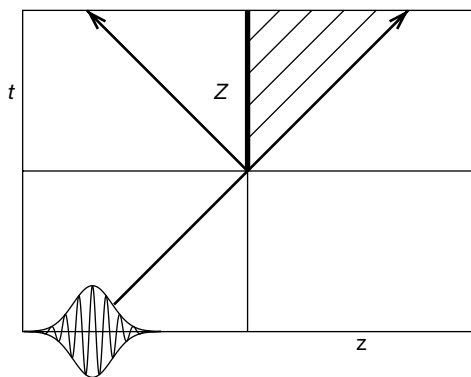


Figure 3 Space–time diagram of a slow-light catastrophe. The figure illustrates the fate of a wave packet $\varphi(t, z)$ that experiences the formation of the horizon Z . Initially, the packet oscillates with positive frequencies in time t , and propagates from the left to the right in space z . The horizon cannot generate negative frequencies in the reflected light, apart from a brief burst that is neglected here. On the left side of Z , we thus regard $\varphi(t, z)$ as analytic¹⁸ in t on the lower half of the complex t plane. Furthermore, $\varphi(t, z)$ is analytic in z on the upper half plane throughout the history of the wave packet, because the process of equation (2) conserves analyticity¹⁸. Yet $\varphi(t, z)$ is not analytic in t on the other side of the horizon, as the solution (equation (8)) indicates. Here waves with negative frequencies are continuously peeling away from the horizon, corresponding to a stationary creation of slow-light quanta.

supplemented with analogous formulas where R and L are interchanged. The vacuum modes of equation (8) contain complex-conjugated w modes with negative frequencies and hence negative energies. This is the decisive sign of particle creation^{19,20}.

Similar to a gravitational collapse², the tuning of the control field towards a parabolic intensity profile triggers a wave catastrophe. In turn, the slow-light quantum field sets out to deplete the control beam, taking energy from it, in an attempt to alter the intensity profile that has caused the catastrophe—but in vain. The control beam continuously replenishes the profile, driving a stationary production of slow-light quantum pairs. The two particles of each pair are created on opposite sides of the horizon, they depart at a snail's pace, accelerate gradually and emerge as detectable photons, like the Hawking radiation^{3,19} of black holes. In contrast to gravitational holes, it is possible to explore the other side of the horizon and measure the non-local correlations²² of the photon pairs. The weight of the negative-frequency component w^* in the vacuum modes of equation (8) gives²⁰ the average photon number per mode:

$$\bar{n} = \frac{1}{(e^{\pi\mu} + e^{-\pi\mu})^2} \quad (9)$$

Maximally 1/4 photons are created on average, which is quite substantial, considering the fact that bright sunlight carries a mere 0.01 photons per mode in the optical range of the Planck spectrum²². Yet the pair production occurs in a narrow frequency window above the critical frequency $\omega_0 + c^2/(8a^2\omega_0)$ which, for realistic experimental parameters^{4,6}, reduces the total photon flux to a few millions of particles per second, a respectable rate. The radiation could perhaps be seen with the naked eye. The characteristic length a of the group-velocity profile (equation (1)) determines the spectral width (equation (9)) of the pair production. The steeper the control-field gradient is, the more quanta are created. Black holes show a similar behaviour³. The smaller the hole is, the larger is the gravity gradient at the horizon and the stronger is the radiation generated³.

Close to the horizon the susceptibility of slow light diverges. Yet infinite susceptibilities tend to be prevented: instead of responding infinitely strongly, optical media become absorptive or nonlinear. According to the physics of EIT illustrated in Fig. 2, the optical nonlinearity of slow light depends on the ratio of the probe and control intensities, I_p and I_c . Equation (6) shows that $I_p \propto |\varphi|^2$ grows linearly in z , whereas I_c is quadratic in z . Therefore, at a critical distance from the horizon the EIT medium becomes nonlinear. A detailed three-dimensional calculation, using the parameters of the experiments of refs 4 and 6, indicates that the nonlinearity sets in before the absorption of the medium becomes important, given a sufficiently steep control-intensity gradient. The Rabi frequency²² of the control light should grow as the distance $|z|$ from the horizon increases, at a rate of at least 10 MHz per wavelength $\lambda_0 = 2\pi/k_0$. In this case, the scale a is about $5 \times 10^3 \lambda_0$.

The quantum radiation of a slow-light catastrophe resembles Hawking radiation, but also exhibits some interesting differences.

Table 1 Quantum catastrophes

Example	Exponent	Average particle number
Hawking radiation	$i\mu$	$\frac{1}{e^{2\pi\mu} - 1}$
Unruh effect		
Schwinger's pair production	$i\mu - 1/2$	$e^{-2\pi\mu}$
Slow light	$i\mu + 1/2$	$\frac{1}{(e^{\pi\mu} - e^{-\pi\mu})^2}$

In each example, a wave develops a singularity with a characteristic exponent. Quantum physics resolves the singularity, and produces particle pairs with a characteristic spectrum (average particle number).

The emitted spectrum (equation (9)) is not planckian, whereas a black hole of Schwarzschild radius r_s appears as a black-body radiator with temperature³ $\hbar c/(4\pi r_s)$. The differences between the two spectra can be traced back to two different classes of wave catastrophes. In both cases¹⁹, waves freeze at an horizon in the form ζ^p with an exponent p of $i\mu + 1/2$ for slow-light media but with an exponent $i\mu$ for black holes where $\mu = 2\pi r_s \omega/c$. Note that Unruh's effect²¹ of radiation seen by an accelerated observer is of the Hawking class as well¹⁹, and so are most of the proposed artificial black holes^{15,23–29}. Schwinger's pair production of charged particles in electrostatic fields³⁰ is accompanied by a subtle wave catastrophe of exponent¹⁹ $i\mu - 1/2$, and leads to a boltzmannian spectrum $\bar{n} = \exp(-2\pi\mu)$. All three catastrophes agree in the limit of large μ , but deviate significantly in the regime of maximal particle production where μ is small (see Table 1). It would be interesting to find out whether more than three types of quantum catastrophe can occur. □

Received 4 June; accepted 19 November 2001.

- Berry, M. V. & Upstill, C. Catastrophe optics: morphologies of caustics and their diffraction patterns. *Prog. Opt.* **XVII**, 257–346 (1980).
- Misner, Ch. W., Thorne, K. S. & Wheeler, J. A. *Gravitation* (Freeman, New York, 1999).
- Hawking, S. M. Black hole explosions? *Nature* **248**, 30–31 (1974).
- Liu, Ch., Dutton, Z., Behroozi, C. H. & Hau, L. V. Observation of coherent optical information storage in an atomic medium using halted light pulses. *Nature* **409**, 490–493 (2001).
- Philips, D. F., Fleischhauer, A., Mair, A., Walsworth, R. L. & Lukin, M. D. Storage of light in atomic vapor. *Phys. Rev. Lett.* **86**, 783–786 (2001).
- Dutton, Z., Budde, M., Slowe, C. & Hau, L. V. Observation of quantum shock waves created with ultra-compressed slow light pulses in a Bose-Einstein condensate. *Science* **293**, 663–668 (2001).
- Born, M. & Wolf, E. *Principles of Optics* (Cambridge Univ. Press, Cambridge, 1999).
- Thom, R. *Stabilité Structurelle et Morphogénèse* (Benjamin, Reading, 1972).
- Poston, T. & Stewart, I. *Catastrophe Theory and its Applications* (Dover, Mineola, 1996).
- Schleich, W. & Scully, M. O. in *New Trends in Atomic Physics* (Les Houches session XXXVIII, Elsevier, Amsterdam, 1984).
- Wang, L. J., Kuzmich, A. & Dogariu, A. Gain-assisted superluminal light propagation. *Nature* **406**, 277–279 (2000).
- Hau, L. V., Harris, S. E., Dutton, Z. & Behroozi, C. H. Light speed reduction to 17 metres per second in an ultracold atomic gas. *Nature* **397**, 594–598 (1999).
- Scully, M. O. & Zubairy, M. S. *Quantum Optics* (Cambridge Univ. Press, Cambridge, 1997).
- Fleischhauer, M. & Lukin, M. D. Dark-state polaritons in electromagnetically-induced transparency. *Phys. Rev. Lett.* **84**, 5094–5097 (2000).
- Leonhardt, U. & Pinewicki, P. Relativistic effects of light in moving media with extremely low group velocity. *Phys. Rev. Lett.* **84**, 822–825 (2000).
- Weinberg, S. *The Quantum Theory of Fields* (Cambridge Univ. Press, Cambridge, 1999).
- Erdélyi, A., Magnus, W., Oberhettinger, F. & Tricomi, F. G. *Higher Transcendental Functions* (McGraw-Hill, New York, 1981).
- Ablowitz, M. J. & Fokas, A. S. *Complex Variables* (Cambridge Univ. Press, Cambridge, 1997).
- Brout, R., Massar, S., Parentani, R. & Spindel, Ph. A primer for black hole quantum physics. *Phys. Rep.* **260**, 329–446 (1995).
- Birrell, N. D. & Davies, P. C. W. *Quantum Fields in Curved Space* (Cambridge Univ. Press, Cambridge, 1982).
- Unruh, W. G. Notes on black-hole evaporation. *Phys. Rev. D* **14**, 870–892 (1976).
- Mandel, L. & Wolf, E. *Optical Coherence and Quantum Optics* (Cambridge Univ. Press, Cambridge, 1995).
- Unruh, W. G. Experimental black-hole evaporation? *Phys. Rev. Lett.* **46**, 1351–1353 (1981).
- Visser, M. Acoustic black holes: horizons, ergospheres and Hawking radiation. *Class. Quantum Grav.* **15**, 1767–1791 (1998).
- Jacobson, T. A. & Volovik, G. E. Event horizons and ergoregions in ³He. *Phys. Rev. D* **58**, 064021–1–064021-7 (1998).
- Volovik, G. E. Simulation of a Panlevé-Gullstrand black hole in a thin ³He-A film. *JETP Lett.* **69**, 705–713 (1999).
- Garay, L. J., Anglin, J. R., Cirac, J. I. & Zoller, P. Sonic analog of gravitational black holes in Bose-Einstein condensates. *Phys. Rev. Lett.* **85**, 4643–4647 (2000).
- Reznik, B. Origin of the thermal radiation in a solid-state analogue of a black hole. *Phys. Rev. D* **62**, 044044-1–044044-7 (2000).
- Chapline, G., Hohlfield, E., Laughlin, R. B. & Santiago, D. I. Quantum phase transitions and the breakdown of classical general relativity. *Phil. Mag. B* **81**, 235–254 (2001).
- Schwinger, J. On gauge invariance and vacuum polarization. *Phys. Rev.* **82**, 664–679 (1951).

Acknowledgements

The Royal Institution discussion meeting on artificial black holes was an inspiration for this work, and I thank the participants and organizers. I acknowledge the support of the ESF programme Cosmology in the Laboratory.

Correspondence and requests for materials should be addressed to the author (e-mail: ulf@st-andrews.ac.uk).

Cyclotron resonance of composite fermions

I. V. Kukushkin^{*†}, J. H. Smet^{*}, K. von Klitzing^{*} & W. Wegscheider^{‡§}

^{*}Max-Planck-Institut für Festkörperforschung, D-70569 Stuttgart, Germany

[†]Institute of Solid State Physics, RAS, Chernogolovka, 142432 Russia

[‡]Walter Schottky Institut, Technische Universität München, D-85748 Garching, Germany

[§]Institut für Experimentelle und Angewandte Physik, Universität Regensburg, D-93040 Regensburg, Germany

It is occasionally possible to interpret strongly interacting many-body systems within a single-particle framework by introducing suitable fictitious entities, or 'quasi-particles'. A notable recent example of the successful application of such an approach is for a two-dimensional electron system that is exposed to a strong perpendicular magnetic field. The conduction properties of the system are governed by electron–electron interactions, which cause the fractional quantum Hall effect¹. Composite fermions, electrons that are dressed with magnetic flux quanta pointing opposite to the applied magnetic field, were identified as apposite quasi-particles that simplify our understanding of the fractional quantum Hall effect^{2,3}. They precess, like electrons, along circular cyclotron orbits, but with a diameter determined by a reduced effective magnetic field^{4–10}. The frequency of their cyclotron motion has hitherto remained enigmatic, as the effective mass is no longer related to the band mass of the original electrons and is entirely generated from electron–electron interactions. Here we demonstrate enhanced absorption of a microwave field in the composite fermion regime, and interpret it as a resonance with the frequency of their circular motion. From this inferred cyclotron resonance, we derive a composite fermion effective mass that varies from 0.7 to 1.2 times that of the electron mass in vacuum as their density is tuned from $0.6 \times 10^{11} \text{ cm}^{-2}$ to $1.2 \times 10^{11} \text{ cm}^{-2}$.

The attachment of two flux quanta (more generally an even number) to composite fermions (CFs) is a natural way to minimize the energy of the two-dimensional electron system (2DES), as the associated vortex expels other electrons from its neighbourhood and decreases the repulsive interaction between the two-dimensional electrons. If two flux quanta per electron penetrate the sample, that is, if the filling factor ν equals 1/2, the external field is effectively compensated and a metallic state of these compound particles emerges⁴. This state can be characterized by a Fermi wavevector and energy. A deviation from exact half-filling results in the appearance of a non-zero effective field B^* , which quantizes the CF motion and discretizes their energy spectrum into Landau levels. In this framework, the fractional quantum Hall effect¹ (FQHE) is a manifestation of this Landau quantization and is equivalent to the integer quantum Hall effect of CFs. A variety of experimental observations^{5–10} can be understood in semiclassical terms of nearly independent CFs.

As the kinetic energy of electrons is entirely quenched when a magnetic field B is applied, the CF cyclotron mass is not a renormalized version of the electron conduction band mass, but originates entirely from Coulomb interactions⁴. The search for the cyclotron resonance of the CFs requires more sophisticated methods than those used to detect the electron cyclotron resonance, because the consequences of Kohn's theorem¹¹ must be avoided. This theorem states that homogeneous radiation in a translationally invariant system can only couple to the centre-of-mass coordinate and cannot excite other internal degrees of freedom. Thus, phenomena originating from electron–electron interactions will not be reflected in the absorption spectrum. One way to bypass the results

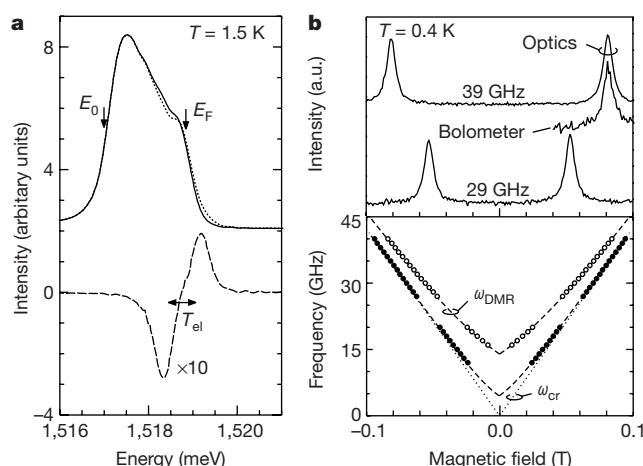


Figure 1 The optical scheme to detect resonant microwave absorption for the electron cyclotron-magnetoplasmon hybrid mode at low magnetic fields. GaAs/Al_{0.3}Ga_{0.7}As heterostructures, containing a single 30-nm-wide quantum well, were used for the investigation. The embedded 2DESs have densities and electron transport mobilities between $0.6\text{--}1.5 \times 10^{11} \text{ cm}^{-2}$ and $3.5 \times 10^6 \text{ cm}^2 \text{ V}^{-1} \text{ s}^{-1}$ respectively. **a**, Luminescence spectrum with (dotted line) and without (solid line) a 50- μW microwave excitation at 18 GHz obtained in a disk-shaped 2DES with a diameter of 1 mm and carrier density $n_s = 5.8 \times 10^{10} \text{ cm}^{-2}$ at a magnetic field $B = 22 \text{ mT}$. The spectra were recorded by using a CCD (charge coupled device) camera, a double-grating spectrometer that provides a spectral resolution of 0.03 meV, and a stabilized semiconductor laser operating at a wavelength of 750 nm and approximately 100 μW of continuous wave power. In the vicinity of the Fermi energy E_F , the spectrum is affected significantly under resonant microwave excitation owing to heating. The dashed line represents the differential

luminescence spectrum. The width of the differential spectrum reflects the increased electron temperature T_{el} . The integration of its absolute value across the entire spectral range yields the microwave absorption amplitude. **b**, Top panel: the microwave absorption amplitude at 29 GHz and 39 GHz as a function of magnetic field obtained by recording differential luminescence spectra as in **a** for 1 mT field increments at $n_s = 1.09 \times 10^{11} \text{ cm}^{-2}$. The peaks, symmetrically arranged around zero field, are identified as the dimensional magnetoplasma-cyclotron hybrid mode. The middle trace shows a conventional bolometer measurement. Bottom panel: resonance position for $n_s = 1.09 \times 10^{11}$ (open circles) and $1.1 \times 10^{10} \text{ cm}^{-2}$ (solid circles) as a function of incident microwave frequency. The intervals 10–20 GHz and 27–40 GHz were covered. The dashed lines represent the theoretical dependence of the hybrid dimensional magnetoplasma-cyclotron resonance. The dotted line corresponds to the cyclotron mode only.

of this theorem is to impose a periodic density modulation to break translational invariance. The non-zero wavevectors defined by the appropriately chosen modulation may then offer access to the cyclotron transitions of CFs, even though they are likely to remain very weak. Therefore, a prerequisite for our studies was the development of an optical detection scheme that boosts the sensitivity to resonant microwave absorption by up to two orders of magnitude in comparison with traditional techniques. Furthermore, we exploited the accidental discovery that microwaves already incident on the sample set up a periodic modulation through the excitation of surface acoustic waves (SAWs).

The 2DES, patterned into a 1-mm-diameter disk¹², was placed near the end of a 16- or 8-mm short-circuited waveguide in the electric field maximum of the microwave excitation inside a ³He cryostat. At a fixed magnetic field B , luminescence spectra with and without microwave excitation were recorded consecutively. The differential luminescence spectrum is obtained by subtracting the spectrum without microwave irradiation from the spectrum obtained under microwave excitation. To improve signal to noise ratio, the same procedure was repeated N times ($N = 2\text{--}20$). Subsequently, we integrated the absolute value of the averaged differential spectrum over the entire spectral range, and called the value of this integral ‘the microwave absorption amplitude’. The same procedure was then repeated for different values of B . To establish the reliability of this scheme, we apply it in Fig. 1 to the well known case of the electron cyclotron resonance $\omega_{\text{cr}} = eB/m^*$, where e is the electron charge and m^* is the effective mass of GaAs (0.067 times the electron mass in vacuum, m_0). Owing to its limited size, the sample also supports a dimensional plasma mode with a frequency ω_p , which depends on both the density n_{2D} and diameter d of the sample, according to $\omega_p^2 = 3\pi^2 n_{2D} e^2 / (2m^* \epsilon_{\text{eff}} d)$. The plasma and cyclotron mode hybridize, and the resulting resonance frequency of the upper dimensional magnetoplasma-cyclotron mode

ω_{DMR} equals $\omega_{\text{cr}}/2 + [\omega_p^2 + (\omega_{\text{cr}}/2)^2]^{1/2}$ (refs 12 and 13). The optical method indeed recovers this mode. A comparison with the theoretical expression for ω_{DMR} yields excellent agreement. No fitting is required, as the density can be independently obtained from the luminescence at higher B values, where Landau levels can be

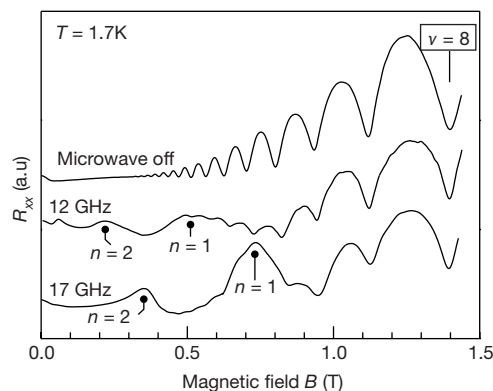


Figure 2 Magnetotransport data in the absence (top curve) and in the presence of 100- μW microwave radiation at 12 GHz (middle curve) and 17 GHz (bottom curve). Curves are offset for clarity. Besides the Shubnikov-de Haas oscillations in the resistance along the current flow (R_{xx}) in a Hall bar geometry¹, additional magneto-resistance oscillations appear under microwave radiation. They are commonly observed in 2DESs on which a static periodic modulation of the density has been imposed. There is no sign of such commensurability effects in the corresponding optical experiments. This can ultimately be traced back to the fact that, contrary to optical quantities, transport is also sensitive to semi-classical phenomena unrelated to changes in the density of states.

resolved. At sufficiently low density, the influence of ω_p on the hybrid mode drops, and at large enough B values the anticipated $\omega_{cr} = eB/m^*$ dependence is recovered. Further details of the electron cyclotron resonance are discussed elsewhere¹². Additional support for the validity of the detection method comes from a comparison with measurements based on the conventional approach using a bolometer (Fig. 1b). Not only do we find the same resonance position, but also the same line shape. The only difference is the improved signal to noise ratio (30–100 times) for the optical scheme, which enables us to observe the electron cyclotron resonance at microwave levels below 10 nW.

Disorder and the finite dimensions of the sample in principle suffice to break translational invariance, as shown by the interaction of the cyclotron with the dimensional plasma mode. However, they provide access to internal degrees of freedom, other than the electron centre of mass motion, at wavevectors that are either poorly defined or too small for appropriate sample sizes. It is therefore desirable to impose a periodic density modulation that introduces larger and well defined wavevectors to circumvent Kohn's theorem.

Transport experiments in the Hall bar geometry showed that it is not necessary to perform additional processing because the microwaves, already incident on the sample, also induce a periodic modulation at sufficiently high power. A clear signature is the appearance of commensurability oscillations in the magneto-resistance owing to the interplay between the B -dependent cyclotron radius of electrons and the length scale of the modulation¹⁴. Examples are displayed in Fig. 2 and resemble the data in ref. 15, where the modulation is produced with the help of SAW transducers. Here, the following model is conceivable. Owing to the piezo-

electric properties of the $\text{Al}_x\text{Ga}_{1-x}\text{As}$ crystal, the radiation is partly transformed into SAWs with opposite momentum, so that both energy and momentum are conserved. Reflection from cleaved boundaries of the sample then produces a standing wave with a periodicity determined by the sound wavelength. Photo-excitation creates a very poorly conducting parallel three-dimensional layer in the Si-doped portion of the AlGaAs barrier, and this may enhance the influence of the standing acoustic waves. Carriers are collected in the nodes and affect the local density of the 2DES. The involvement of sound waves can be deduced from transport data, because from the minima we expect the modulation period to be approximately 200 and 250 nm for frequencies of 17 and 12 GHz respectively. The ratio of this period to the sound wavelength at these frequencies is 1.12 and 1.15.

Figure 3a depicts the amplitude of the microwave absorption up to high values of B . Apart from the strong dimensional magneto-plasma-cyclotron resonance signal at low magnetic field discussed above, several peaks, that scale with a variation of the density, emerge near filling factors 1, 1/2 and 1/3. The position of the peaks associated with $\nu = 1$ and 1/3 remains fixed when tuning the microwave frequency, and are ascribed to heating induced by non-resonant absorption of microwave power. In contrast, the weak maxima surrounding filling 1/2 readily respond to a change in frequency, as illustrated in Fig. 3b. They are symmetrically arranged around filling 1/2, and their splitting increases with frequency. The dependence on B is summarized in Fig. 3c for two densities. To emphasize the symmetry, B^* was chosen as the abscissa. The linear relationship between frequency and field extrapolates to zero at vanishing B^* . We do not expect a deviation at small B^* owing to a plasma-like contribution, as seen in Fig. 1c. Excitations for the 1/3, 2/5, 3/7 and other fractional quantum Hall states exhibit in numerical simulations no magnetoplasmon-like linear contribution to the dispersion at small values of momentum, k (ref. 16). We suggest that the resonance observed in Fig. 3 is the long-sought-for cyclotron resonance of CFs. An alternative interpretation of the observed features in terms of geometric resonances (GRs), as they occur in transport at low fields due to the density modulation (Fig. 2), would contradict several observations: (1) in the optical data, only the electron cyclotron resonance peak is observed—and no commensurability effects. Unlike optical quantities, transport is also sensitive to semi-classical phenomena unrelated to changes in the density of states. We therefore anticipate no commensurability

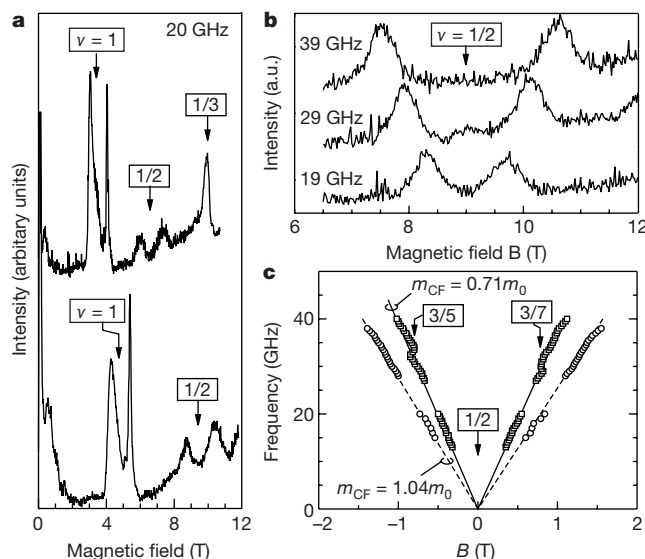


Figure 3 Microwave absorption amplitude and peak position for high magnetic fields in the samples with a disk-shaped mesa at $T = 0.4$ K. **a**, Microwave absorption amplitude at high magnetic fields for $n_s = 0.81 \times 10^{11}$ and $1.15 \times 10^{11} \text{ cm}^{-2}$ and frequency of 20 GHz. The response near $\nu = 1$ and 1/3 does not shift with frequency. **b**, Microwave absorption amplitude in the vicinity $\nu = 1/2$ (20 mT step size) at three different frequencies and $n_s = 1.09 \times 10^{11} \text{ cm}^{-2}$. The peak values are nearly two orders of magnitude weaker and considerably wider (about 30–50 times) than those due to the electron cyclotron resonance. **c**, Position of the CF cyclotron mode as a function of the effective magnetic field $B^* = B - 2n_s\phi_0$ (ϕ_0 is the elementary flux quantum) for $n_s = 1.09 \times 10^{11} \text{ cm}^{-2}$ (circles) and $n_s = 0.59 \times 10^{11} \text{ cm}^{-2}$ (squares). The corresponding CF effective masses m_{CF} equal $1.04 m_0$ and $0.71 m_0$ respectively. The resonances remain visible even if the 2DES condenses in fractional quantum Hall states at filling factors 3/5 and 3/7.

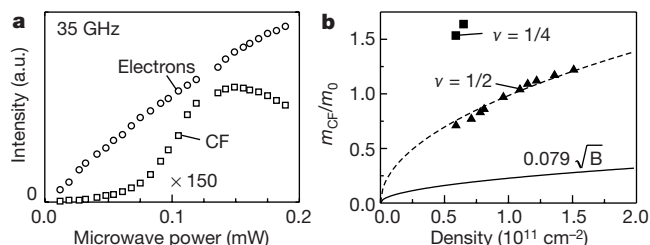


Figure 4 Cyclotron resonance amplitude as a function of incident microwave power and cyclotron mass of CFs at $\nu = 1/2$ and $1/4$ for various carrier densities. **a**, Incident microwave power dependence of the amplitude at the cyclotron resonance (circles for electrons, squares for CFs). **b**, Dependence of the CF effective mass near $\nu = 1/2$ on the carrier density n_s (solid triangles). The mass for $\nu < 1/2$ is systematically a few per cent heavier than for $\nu > 1/2$. The dashed line is a square-root fit to the data. The solid curve is the prediction from theory reported in ref. 17. Analogous resonance peaks have also been detected around $\nu = 1/4$. The corresponding effective mass values are indicated as solid squares for two different densities. The large discrepancy in the extracted values of the effective masses for $\nu = 1/2$ and $\nu = 1/4$ allows us to exclude the possibility that the resonant microwave absorption peaks originate from a commensurability effect.

features near $\nu = 1/2$ in optics. (2) Even if the 2DES condenses in a FQHE state and the chemical potential is located in a gap, the resonance peaks surrounding $\nu = 1/2$ survive (Fig. 3c). Commensurability effects should not be observable in this regime. (3) Similar resonance peaks were also detected for the higher-order CFs around $\nu = 1/4$. As this CF metallic state is characterized by the same Fermi wavevector, GRs should occur at the same distance from $\nu = 1/4$ as they do at $\nu = 1/2$. The observed peaks are located at different positions, making it difficult to support a commensurability situation (see below, and Fig. 4b).

In contrast to the electron case, the intensity of the CF cyclotron resonance is a strong nonlinear function of microwave power (Fig. 4a). Moreover, its observability only at high power correlates with the first appearance of commensurability oscillations. The drop in intensity at even higher power is probably caused by heating. The intensity diminishes to zero at temperatures above 0.7 K, whereas the electron resonance persists up to $T > 2$ K. The slope of the CF cyclotron frequency as a function of B^* in Fig. 3c defines the cyclotron mass $m_{\text{cf}}^{\text{cf}}$. This mass is set by the electron–electron interaction scale, so that a square-root dependence on density or magnetic field is predicted from a straightforward dimensional analysis⁴. Numerical calculations predict $m_{\text{cf}}^{\text{cf}}/m_0 = 0.079(B)^{1/2}$ for an ideal 2DES (here B is in tesla), not including Landau level mixing or finite width contributions¹⁷. The data, shown in Fig. 4b, confirm qualitatively the marked enhancement in comparison with the electron mass (more than 10 times); however, a fit to the square-root dependence requires a prefactor that is four times larger. Previously reported mass values based on measurements of the activation energy gap^{18,19} must be distinguished from the cyclotron mass. The former corresponds to the limit of infinite momentum, whereas here k approaches zero. Moreover, activation gaps can only be extracted at well developed fractional quantum Hall states, and their accurate determination suffers from disorder-induced broadening. These and other limitations have been discussed in ref. 19. The present technique can be performed at arbitrary filling factors. □

Received 27 April; accepted 14 November 2001.

1. Das Sarma, S. & Pinczuk, A. (eds) *Perspectives on Quantum Hall Effects* (Wiley, New York, 1996).
2. Jain, J. K. The composite fermion: a quantum particle and its quantum fluids. *Physics Today* **53** (4), 39–45 (2000).
3. Jain, J. K. Composite fermion approach for the fractional quantum Hall effect. *Phys. Rev. Lett.* **63**, 199–202 (1989).
4. Halperin, B. I., Lee, P. A. & Read, N. Theory of the half-filled Landau level. *Phys. Rev. B* **47**, 7312–7343 (1993).
5. Heinonen, O. (ed.) *Composite Fermions* (World Scientific, Singapore, 1998).
6. Du, R. R., Stormer, H. L., Tsui, D. C., Pfeiffer, L. N. & West, K. W. Experimental evidence for new particles in the fractional quantum Hall effect. *Phys. Rev. Lett.* **70**, 2944–2947 (1993).
7. Du, R. R. *et al.* Fractional quantum Hall effect around $\nu = 3/2$: composite fermions with a spin. *Phys. Rev. Lett.* **75**, 3926–3929 (1995).
8. Willett, R. L., Ruel, R. R., West, K. W. & Pfeiffer, L. N. Experimental demonstration of a Fermi surface at one-half filling of the lowest Landau level. *Phys. Rev. Lett.* **71**, 3846–3849 (1993).
9. Goldman, V. J., Su, B. & Jain, J. K. Detection of composite fermions by magnetic focusing. *Phys. Rev. Lett.* **72**, 2065–2068 (1994).
10. Smet, J. H., Weiss, D., Blick, R. H., Lütjering, G. & von Klitzing, K. Magnetic focusing of composite fermions through arrays of cavities. *Phys. Rev. Lett.* **77**, 2272–2275 (1996).
11. Kohn, W. Cyclotron resonance and de Haas-van Alphen oscillations of an interacting electron gas. *Phys. Rev.* **123**, 1242–1244 (1961).
12. Gubarev, S. I. *et al.* Screening of excitonic states by low-density 2D charge carriers in GaAs/AlGaAs quantum wells. *JETP Lett.* **72**, 324–328 (2000).
13. Allen, S. J., Störmer, H. L. & Hwang, J. C. M. Dimensional resonance of the two-dimensional electron gas in selectively doped GaAs/AlGaAs heterostructures. *Phys. Rev.* **28**, 4875–4877 (1983).
14. Gerhardt, R. R., Weiss, D. & von Klitzing, K. Novel magnetoresistance oscillations in a periodically modulated two-dimensional electron gas. *Phys. Rev. Lett.* **62**, 1173–1176 (1989).
15. Shilton, J. M. *et al.* Effect of spatial dispersion on acoustoelectric current in a high mobility two-dimensional electron gas. *Phys. Rev. B* **51**, 14770–14773 (1995).
16. Jain, J. K. & Kamilla, R. K. In *Composite Fermions* (ed. Heinonen, O.) 1–80 (World Scientific, Singapore, 1998).
17. Park, K. & Jain, J. K. Phase diagram of the spin polarization of composite fermions and a new effective mass. *Phys. Rev. Lett.* **80**, 4237–4240 (1998).
18. Du, R. R., Stormer, H. L., Tsui, D. C., Pfeiffer, L. N. & West, K. W. *Phys. Rev. Lett.* **70**, 2944–2947 (1993).
19. Willett, R. L. In *Composite Fermions* (ed. Heinonen, O.) 349–431 (World Scientific, Singapore, 1998).

Acknowledgements

We acknowledge the Volkswagen Stiftung, the Russian Fund of Fundamental Research, INTAS, the German Ministry of Science and Education and the German Physical Society for their support.

Competing interests statement

The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to J.H.S. (e-mail: jhsmet@klizix.mpi-stuttgart.mpg.de).

Imaging the granular structure of high- T_c superconductivity in underdoped $\text{Bi}_2\text{Sr}_2\text{CaCu}_2\text{O}_{8+\delta}$

K. M. Lang*, V. Madhavan*, J. E. Hoffman*, E. W. Hudson*†‡, H. Eisaki§, S. Uchida§ & J. C. Davis*‡

* Department of Physics, University of California, Berkeley, California 94720, USA

† Department of Physics, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139-4301, USA

‡ Materials Science Division, Lawrence Berkeley National Laboratory, Berkeley, California 94720, USA

§ Department of Superconductivity, University of Tokyo, Yayoi, 2-11-16 Bunkyo, Tokyo 113-8656, Japan

|| Department of Applied Physics, Stanford University, Stanford, California 94205-4060, USA

Granular superconductivity occurs when microscopic superconducting grains are separated by non-superconducting regions; Josephson tunnelling between the grains establishes the macroscopic superconducting state¹. Although crystals of the copper oxide high-transition-temperature (high- T_c) superconductors are not granular in a structural sense, theory suggests that at low levels of hole doping the holes can become concentrated at certain locations resulting in hole-rich superconducting domains^{2–5}. Granular superconductivity arising from tunnelling between such domains would represent a new view of the underdoped copper oxide superconductors. Here we report scanning tunnelling microscope studies of underdoped $\text{Bi}_2\text{Sr}_2\text{CaCu}_2\text{O}_{8+\delta}$ that reveal an apparent segregation of the electronic structure into superconducting domains that are ~ 3 nm in size (and local energy gap < 50 meV), located in an electronically distinct background. We used scattering resonances at Ni impurity atoms⁶ as ‘markers’ for local superconductivity^{7–9}; no Ni resonances were detected in any region where the local energy gap $\Delta > 50 \pm 2.5$ meV. These observations suggest that underdoped $\text{Bi}_2\text{Sr}_2\text{CaCu}_2\text{O}_{8+\delta}$ is a mixture of two different short-range electronic orders with the long-range characteristics of a granular superconductor.

Evidence from scanning tunnelling microscopy (STM) for nanoscale spatial variations in the electronic characteristics of $\text{Bi}_2\text{Sr}_2\text{CaCu}_2\text{O}_{8+\delta}$ (Bi-2212) has been steadily accumulating^{10–12}. Recent advances include observations of nanoscale regions with ‘superconducting’ spectra coexisting with regions characterized by ‘pseudogap-like’ spectra^{13,14}, and the observation of spatial inhomogeneities in the gap magnitude that correlate with those in integrated differential tunnelling conductance¹⁵. Several theoretical models have emerged in response to these data^{16–22}. A frequent theme among these models is that strong electrostatic-potential variations are experienced by the itinerant holes. Such potential fluctuations could be caused by frustrated electronic phase separation^{2–5}, by unscreened dopant atoms^{15,19–21}, or by some other

unrelated disorder^{16–18,22}. In any of these cases, the holes would redistribute themselves to minimize their total energy in the disordered potential landscape. At low hole densities, some regions would therefore remain doped while others would be relatively undoped. Thus, two types of electronic order—superconducting and non-superconducting respectively—would coexist in different nanoscale regions of the same crystal^{16–19}. If Josephson tunnelling were to couple the superconducting domains, the crystal would exhibit granular superconductivity^{13,23}.

These predictions can be explored by STM, as it can access real-space electronic structure with atomic resolution. Here we report high-resolution STM studies of underdoped Bi-2212. The samples are single crystals grown by the floating zone method. The ‘as-

grown’ samples have a hole-dopant level $p \approx 0.18 \pm 0.02$ while the underdoped samples are oxygen-depleted with a superconducting transition temperature T_c of 79 K and $p \approx 0.14 \pm 0.02$. They are cleaved in cryogenic ultrahigh vacuum at $T < 30$ K and, when inserted into the STM head, show atomic resolution on the BiO surface plane. Our primary data sets consist of differential tunnelling conductance spectra (dI/dV versus V) measured at all locations in a given field of view. These ‘spectral surveys’ are essentially detailed spatial maps of the local density of electronic states (LDOS) as a function of energy E , because $\text{LDOS}(E) \propto dI/dV(V)$ where V is the sample bias, e is the electron charge, and $E = eV$.

Spectral surveys can be analysed to yield maps of the energy gap as a function of location (‘gap maps’) by defining Δ for each dI/dV spectrum as:

$$\Delta \equiv \frac{\Delta_+ - \Delta_-}{2} \quad (1)$$

Here $\Delta_{+(-)}$ is the energy of the first peak above (below) the Fermi level. Peaks from scattering resonances are disregarded, and Δ does not necessarily represent a superconducting energy gap.

In Fig. 1 we show gap maps extracted from two such spectral surveys, each measured on an area of $560 \text{ \AA} \times 560 \text{ \AA}$. They use an identical colour scale. Figure 1a is typical of gap maps on underdoped Bi-2212, while Fig. 1b is typical of the slightly overdoped ‘as-grown’ samples. They are very different in appearance. The as-grown gap map is dominated by large interconnected areas of low Δ (red and yellow), interspersed with filamentary high- Δ areas (blue and black). This is in contrast to the underdoped sample in Fig. 1a in which the compact, spatially distinct, low- Δ regions (red and yellow) are surrounded by interconnected high- Δ regions (blue and black). For the as-grown sample shown, the mean gap value is $\bar{\Delta} = 35.6 \text{ meV}$ with a standard deviation $\sigma = 7.7 \text{ meV}$, while for the underdoped sample shown $\bar{\Delta} = 50 \text{ meV}$ and $\sigma = 8.6 \text{ meV}$. Were these gap maps the only information available, one might suspect that reducing the oxygen doping has merely shifted $\bar{\Delta}$. But as we show below using both additional information from the spectral surveys and new experimental techniques, the differences between as-grown and underdoped samples are more profound.

Even though the spectral surveys yielding Fig. 1 had spatial resolution of $\approx 4 \text{ \AA}$, the spectra can show strong variations from one pixel to the next. Therefore, an accurate description of the electronic structure requires higher-spatial-resolution spectral surveys. Figure 2a shows a typical gap map from such a high-resolution spectral survey acquired with 128×128 pixels on the area ($147 \text{ \AA} \times 147 \text{ \AA}$) indicated by the white box in Fig. 1a. Figure 2b is a map of the gap-edge peak amplitude, $G(\Delta_-)$, from the same spectral survey. In a conventional superconductor this would be a map of the coherence-peak heights. We use the same colour bar to represent Δ in Fig. 2a and $G(\Delta)$ in Fig. 2b, to illustrate the spatial correlations between these two different observables. In these figures we see what appear to be compact, almost circular, domains. They are characterized by a low value of Δ , and a $G(\Delta)$ that rises rapidly from the domain edge to reach a maximum at the domain centre. We will refer to these regions as α -domains. As an example, a single α -domain is identified inside the white circle on Fig. 2a and b. Different α -domains have a different characteristic values of Δ , and multiple α -domains are clumped together at some locations. Intervening between the α -domains are percolative regions characterized by high Δ and low, almost-constant $G(\Delta)$. We will refer to these spectroscopically distinct areas as β -regions.

Figure 2c and d shows the evolution of Δ and $G(\Delta)$, respectively, with distance along the white line in Fig. 2a and b. Within each α -domain, the value of Δ varies little, while $G(\Delta)$ rises rapidly from the perimeter reaching a maximum in the centre. In the surrounding interconnected β -regions, Δ varies slowly while $G(\Delta)$ is low and almost constant. Figure 2 focuses on the typical evolution of Δ and $G(\Delta)$, but these parameters represent only a small subset of the

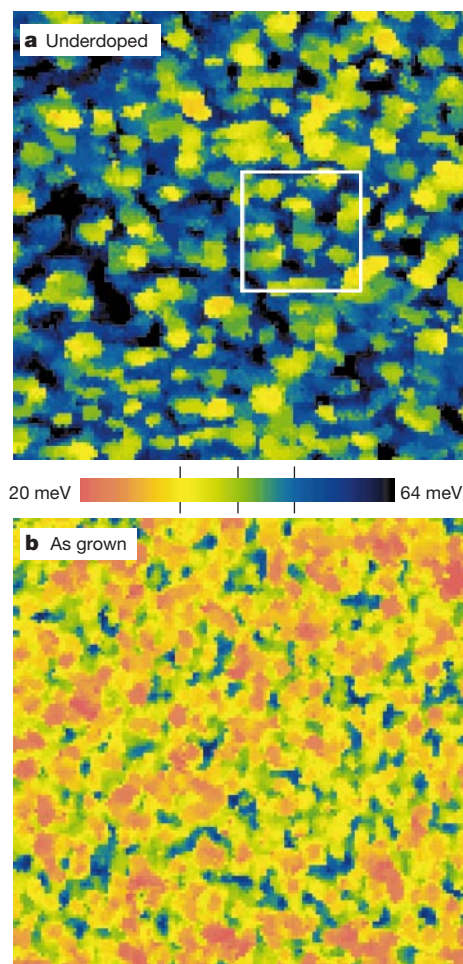


Figure 1 Typical gap maps from underdoped and as-grown Bi-2212 samples, each showing an area of $560 \text{ \AA} \times 560 \text{ \AA}$. We use an identical colour scale to represent Δ in these two panels for ease of comparison. The colour scale spans the range 20–64 meV, and ticks on the scale are placed at 34 meV, 42 meV and 50 meV. These gap maps were calculated from two 128×128 pixel spectral surveys. **a**, Gap map of an oxygen-underdoped Bi-2212 sample with $T_c = 79$ K. This field of view contains approximately 130 compact low- Δ domains which are visible as orange, yellow and green islands against the interconnected blue and black background which are the high- Δ regions. **b**, Gap map of an as-grown Bi-2212 sample. The interconnected red, orange and yellow low- Δ regions completely dominate the image. Embedded in this environment are filamentary blue and black high- Δ regions. Note that although this sample contains a 0.5% substitution of Ni for the Cu atoms in the CuO_2 plane, the 102 impurity resonances in this field of view affect less than 15% of the area, and furthermore, the local value of Δ does not vary near the Ni impurities⁶. Both spectral surveys were acquired with constant-current normalization in cryogenic ultrahigh vacuum at 4.2 K with tunnel junction resistance of $1 \text{ G}\Omega$ set at $V = -100 \text{ mV}$.

information in a spectral survey. To see directly how the electronic phenomena reported here are manifest in the raw data, we show in Fig. 3 the unprocessed dI/dV spectra measured along the white line in Fig. 2a and b. The spectral evolution with passage through three α -domains (red) and the intervening β -regions (blue) can be seen.

Figures 2 and 3 show how, on the basis of two observables (Δ and $G(\Delta)$), the α -domains and β -regions appear spectroscopically distinct. We refer to this situation as electronic segregation. It becomes apparent only in underdoped samples, because $\Delta < 50$ meV for the α -domains and these samples have $\Delta > 50$ meV for $\sim 50\%$ of their area, while the as-grown samples studied have $\Delta > 50$ meV for only $\sim 10\%$ of theirs. This apparent electronic segregation motivates a new picture of the effects of reducing oxygen doping in Bi-2212. The α -domains have characteristics usually associated with superconductivity—for example, sharp gap-edge peaks indicating well defined Bogoliubov quasiparticles and an increasing areal density associated with rising T_c below optimal doping. In contrast, the β -region spectra exhibit characteristics often associated with the pseudogap phase (although we cannot definitively identify these regions as such). Consequently, the emerging picture is reminiscent

of a granular superconductor with coupled superconducting grains embedded in a distinct electronic environment.

As STM is a surface probe, we next consider whether the observed phenomena represent a property of bulk underdoped Bi-2212. Evidence for strong nanoscale variations in the bulk electronic properties includes: (1) a low-temperature quasiparticle lifetime at least an order of magnitude shorter²⁴ in Bi-2212 than in $\text{YBa}_2\text{Cu}_3\text{O}_{7-\delta}$, (2) optical conductivity measurements demonstrating a finite low-temperature σ_1 (ref. 24) that can be explained by strong nanoscale variations in local superfluid density²⁵, (3) heat-capacity ($\equiv C$) measurements showing that with underdoping, $\gamma \equiv C/T$ falls rapidly and the temperature range of the transition widens dramatically²⁶, (4) NMR studies showing very broad line-widths and additional spin susceptibility at low temperature indicative of strong electronic disorder²⁷, (5) inelastic neutron scattering line-widths that are significantly wider in Bi-2212 than in other copper oxides²⁸, (6) ARPES studies showing similar evolution of spectra with deliberately introduced bulk disorder as with underdoping²⁹, and (7) interlayer tunnelling experiments demonstrating the co-existence of superconducting and non-supercon-

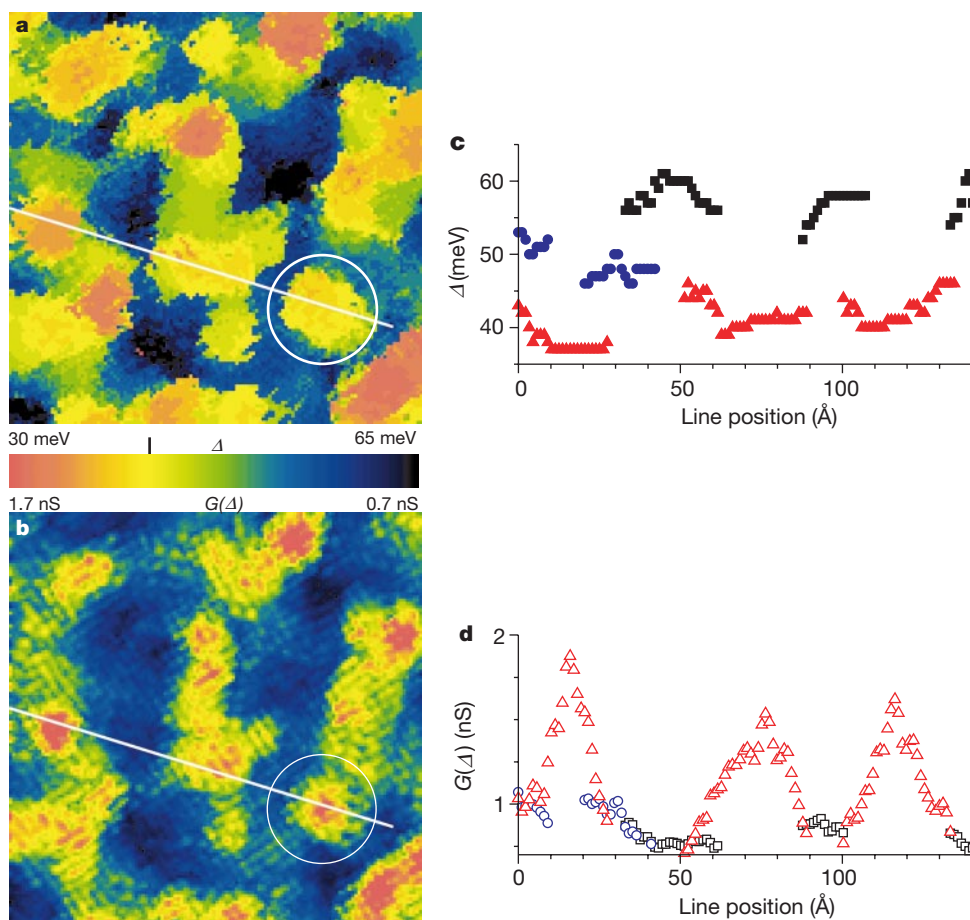


Figure 2 Data obtained at high spatial resolution, showing the typical spatial interrelationship of Δ and $G(\Delta)$ for underdoped Bi-2212. These data were generated from a single 128×128 pixel spectral survey which was taken on the $147 \text{ \AA} \times 147 \text{ \AA}$ region inside the white box in Fig. 1a. This spectral survey was taken several weeks after the survey used to calculate Fig. 1a, during which the sample was kept continuously at 4.2 K. **a**, High-resolution gap map revealing approximately 12 α -domains embedded in the percolative β -like background. **b**, Map of $G(\Delta)$ at the same location as **a**. Here, to reveal the spatial correlations between Δ and $G(\Delta)$, we use the same colour bar as in **a** but here the scale is inverted and represents $G(\Delta)$. Note that $G(\Delta)$ is defined as the gap-edge peak at negative bias throughout this Letter. In **c** we show the spatial evolution of Δ and in **d** of $G(\Delta)$ along the trajectory of the white line seen in **a** and **b**. Figures 2 and 3 together

illustrate the properties of the two distinct types of regions. In terms of Δ , the α -domains are defined by a distinct value of Δ which is ≤ 50 meV and which is constant to 5% within a $13\text{-\AA}$ radius of the centre of each domain, while the β -regions are defined by $\Delta \geq 35$ meV. There exists an overlap range of Δ ($35 \text{ meV} \leq \Delta \leq 50 \text{ meV}$) in which categorization must rely on parameters in addition to Δ . The tick mark on the colour scale is at 42 meV at the centre of this overlap range. In terms of $G(\Delta)$, the α -domains are defined by $G(\Delta)$ at the centre of the domain being $\sim 66\%$ ($\sigma = 23\%$) above the average β -region $G(\Delta)$, which itself has $\sigma = 15\%$. Similar conclusions are derived from all equivalent maps from other underdoped spectral surveys. The spectral survey was acquired with constant current normalization at 4.2 K with tunnel junction resistance of $1 \text{ G}\Omega$ set at $V = -200 \text{ mV}$.

ducting energy gaps^{30,31}. Results from these different techniques would be consistent with each other and with STM data if nanoscale variations exist in the properties of bulk Bi-2212, and if they are amplified by underdoping.

Although the data in Figs 1–3 suggest granular superconductivity, single-particle excitation spectra cannot definitively distinguish between such superconductivity and a highly disordered single-phase superconductor. However, a novel atomic-scale technique, designed to locally distinguish superconducting from non-superconducting regions by using quasiparticle scattering resonances at Ni impurity atoms, has recently been proposed^{7–9}. A scattering resonance originates from interactions between quasiparticles and an impurity atom, and appears as an additional peak (or peaks) in the excitation spectrum near the impurity. Ni scattering resonances are particle–hole symmetric in the superconducting state⁶, are predicted to lose this symmetry if an electronic ‘pseudogap’ is present^{8,9}, and might disappear completely if conventional quasiparticles are non-existent⁷. Thus, study of Ni scattering resonances as a function of local Δ should, in theory, allow the discrimination of superconducting from non-superconducting regions.

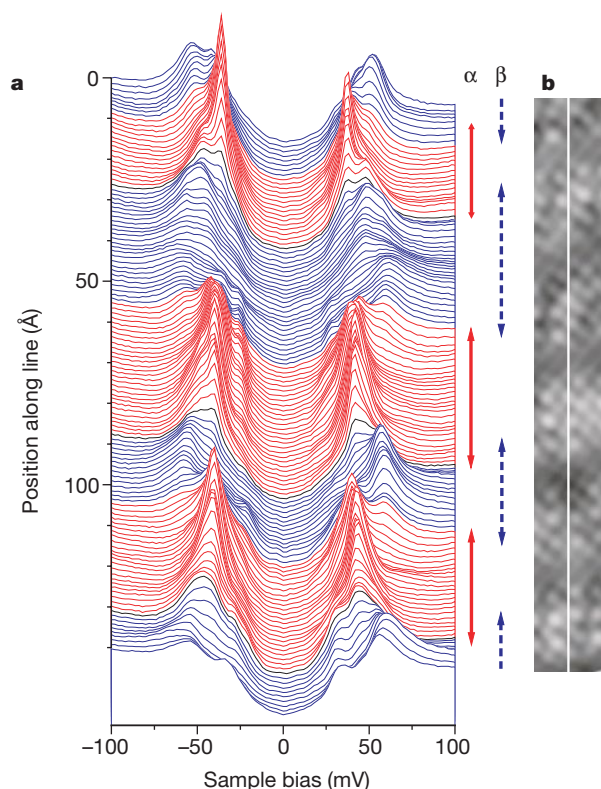


Figure 3 Typical series of dI/dV spectra illustrating how the two distinct types of regions are apparent in the raw data. **a**, These unprocessed spectra were extracted from the same spectral survey used to create Fig. 2, along the white line shown in Fig. 2a and b. The spectra are separated by 1.1 Å and span a total distance of 140 Å as represented by their vertical offset. The α -domain spectra (red) have low gap magnitudes and sharp gap-edge peaks whose amplitude is low at the edges of the domain and rises to a maximum in the centre. The β -region spectra have high gap magnitude and very broad gap-edge peaks whose amplitude is relatively low and constant. Additional features seen in this figure are the spectra that lie on the borders between neighbouring distinct regions. These spectra can show four peaks, where each pair of peaks corresponds to the gap value of the neighbouring regions. For the maps shown in Figs 1 and 2, the peak with the overall maximum value of $G(eV)$ is used to determine Δ and $G(\Delta)$. In Fig. 2c and d, if four clear peaks can be identified in the dI/dV spectrum, then both values of Δ and $G(\Delta)$ at that location are plotted. **b**, Surface topography with the trajectory along which both the spectra in **a**, and the traces of Fig. 2c and d, were measured, shown as a white line. The atomic resolution associated with the spectral surveys is clearly evident.

We apply this technique by carrying out spectral surveys on as-grown Ni-doped Bi-2212 samples. Figure 4 summarizes results from two such experiments. Figure 4a shows the positions of the Ni scattering resonances (which are identified by their strongest resonance peak near +18 meV; ref. 6) superimposed on the gap map derived from the same spectral survey. The colour scale is such that regions with $\Delta < 50$ meV (light grey) are distinguished from those with $\Delta > 50$ meV (dark grey). No Ni scattering resonances are observed in any region where $\Delta > 50 \pm 2.5$ meV. To clarify the correlation between Ni resonances and Δ , we plot in Fig. 4b a combined histogram (shown red) of observed Ni resonances versus local Δ from two independent spectral surveys on different crystals. The combined histogram of Δ from the same two spectral surveys is shown in grey. Although for $\Delta < 35$ meV the two distributions are similar in shape, above $\Delta \approx 35$ meV they rapidly diverge and the Ni-resonance distribution reaches zero by 50 meV.

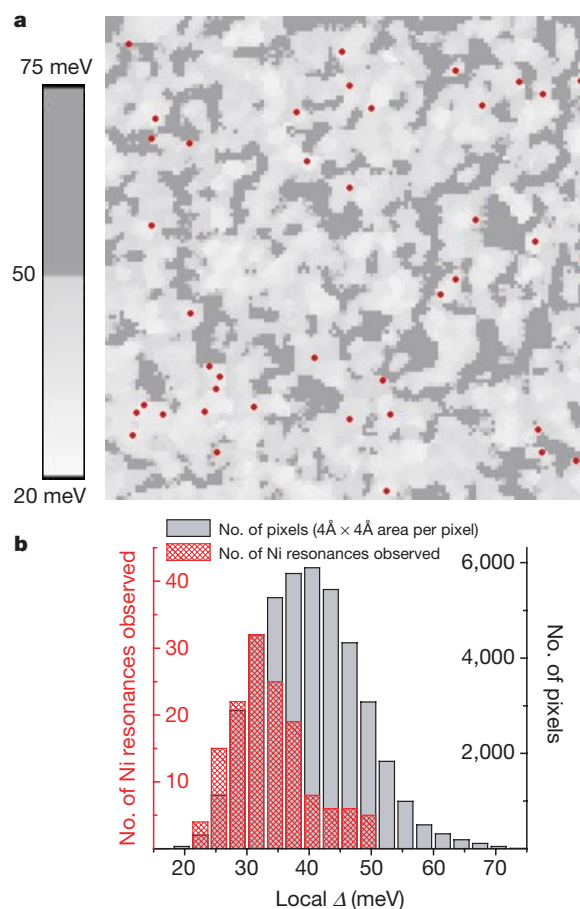


Figure 4 Analysis of the relationship between Ni scattering resonances and local Δ . Spectral surveys were acquired at 4.2 K on two as-grown Ni-doped Bi-2212 samples, both with tunnel junction resistance of 1 GΩ set at $V = -100$ mV. Using the +18 mV LDOS image of each survey, the locations of Ni resonances were identified by the strong resonance peak at +18 mV in their dI/dV spectra⁶. For $\approx 30\%$ of these resonances, the full particle–hole symmetric signature (spatial and spectroscopic) of a Ni resonance⁶ was directly confirmed. The location of the Ni resonances can be correlated with the local value of Δ extracted from these same two surveys. **a**, The location of the Ni resonances (red dots) superimposed on the simultaneously acquired gap map of an area $600 \text{ Å} \times 600 \text{ Å}$. No Ni resonances are observed in regions where $\Delta > 50$ meV. **b**, Histograms derived from a combined analysis of the above spectral survey on 0.2% Ni-doped Bi-2212, and an independent $568 \text{ Å} \times 568 \text{ Å}$ survey on 0.5% Ni-doped Bi-2212. For each Ni resonance, the local Δ is determined by the spatial average of Δ over a 13 Å square region centred on the impurity. Using this information, we plot in red a combined histogram of observed Ni resonances versus Δ from both samples. The combined histogram of Δ from the same two surveys is shown in grey.

There are several possible explanations for the Ni-resonance distribution. First, the amplitude of the 18-meV resonance peak might fall with rising Δ and disappear near $\Delta \approx 50$ meV. However, measurement of the peak heights shows them to be almost independent of local Δ . Second, there might be statistical fluctuations such that no Ni atoms reside in regions where $\Delta > 50$ meV. However, if the Ni atoms are distributed randomly, and if all regions can support quasiparticle scattering resonances, our non-observation (in two independent experiments on different crystals) of Ni resonances in regions with $\Delta > 50$ meV has a combined probability of no more than 3×10^{-5} . Therefore this explanation also appears to be ruled out. A third explanation could be that the Ni atoms somehow 'seed' nanoscale regions (perhaps by attracting the dopant oxygen atoms or the holes), influencing them to develop into superconducting domains with $\Delta < 50$ meV. This model seems unlikely because, in one Zn-doped sample studied, Zn impurity resonances disappear in a statistically similar fashion near $\Delta \sim 50$ meV—but it cannot be ruled out. Nevertheless, impurity atoms are clearly not necessary to create α -domains as they exist in samples with no impurity atoms (see Figs 1a, 2 and 3).

A final hypothesis is that Ni impurity atoms are in fact physically present in regions with $\Delta > 50$ meV, but that they do not create scattering resonances because these regions represent an electronically distinct phase. If Ni atoms are randomly distributed, and if the locations of particle-hole symmetric Ni resonances indicate the local existence of superconductivity, then the distribution of superconducting regions has a similar shape to the red histogram of Fig. 4b. The picture would then be of purely superconducting regions when $\Delta \lesssim 35$ meV, a mixture of two different electronic orders when $35 \text{ meV} \lesssim \Delta \lesssim 50$ meV, and an unidentified second phase (possibly the pseudogap) when $\Delta \gtrsim 50$ meV. The data in Figs 1–3 corroborate this picture, particularly because the energy where the superconducting α -domains disappear is very close to the energy where the Ni resonances disappear. Therefore, although we cannot distinguish between the possible microscopic mechanisms^{2,4,5,15–22} for the phenomena reported here, the data all suggest that underdoped Bi-2212 is a granular superconductor. This provides a new and unconventional context in which to view the underdoped copper oxides. □

Received 28 August; accepted 4 December 2001.

- Simánek, E. *Inhomogeneous Superconductors: Granular and Quantum Effects* (Oxford Univ. Press, New York, 1994).
- Gor'kov, L. P. & Sokol, A. V. Phase stratification of an electron liquid in the new superconductors. *JETP Lett.* **46**, 420–423 (1987).
- Zaenen, J. & Gunnarsson, O. Charge magnetic domain lines and the magnetism of high- T_c oxides. *Phys. Rev. B* **40**, 7391–7394 (1989).
- Emery, V. J., Kivelson, S. A. & Lin, H. Q. Phase separation in the t-J model. *Phys. Rev. Lett.* **64**, 475–478 (1990).
- Emery, V. J. & Kivelson, S. A. Frustrated electronic phase separation and high-temperature superconductors. *Physica C* **209**, 597–621 (1993).
- Hudson, E. W. *et al.* Interplay of magnetism and high- T_c superconductivity at individual Ni impurity atoms in $\text{Bi}_2\text{Sr}_2\text{CaCu}_2\text{O}_{8+x}$. *Nature* **411**, 920–924 (2001).
- Flatté, M. E. Nickel probes superconductivity. *Nature* **411**, 901–903 (2001).
- Flatté, M. E. Quasiparticle resonant states as a probe of short-range electronic structure and Andreev coherence. *Phys. Rev. B* **61**, R14920–R14923 (2000).
- Kruis, H. V., Martin, I. & Balatsky, A. V. Impurity-induced resonant state in a pseudogap state of a high- T_c superconductor. *Phys. Rev. B* **64**, 054501–1–054501–4 (2001).
- Liu, J.-X., Wan, J.-C., Goldman, A. M., Chang, Y. C. & Jiang, P. Z. Features of the density of states of high- T_c superconductors probed by vacuum tunneling. *Phys. Rev. Lett.* **67**, 2195–2198 (1991).
- Chang, A., Rong, Z. Y., Ivanchenko, Y. M., Lu, F. & Wolf, E. L. Observation of large tunneling-conductance variations in direct mapping of the energy gap of single-crystal $\text{Bi}_2\text{Sr}_2\text{CaCu}_2\text{O}_{8-x}$. *Phys. Rev. B* **46**, 5692–5698 (1992).
- Cren, T. *et al.* Influence of disorder on the local density of states in high- T_c superconducting thin films. *Phys. Rev. Lett.* **84**, 147–150 (2000).
- Howald, C., Fournier, P. & Kapitulnik, A. Inherent inhomogeneities in tunneling spectra of $\text{Bi}_2\text{Sr}_2\text{CaCu}_2\text{O}_{8-x}$ crystals in the superconducting state. *Phys. Rev. B* **64**, 100504–1–100504–4 (2001).
- Cren, T., Roditchev, D., Sacks, W. & Klein, J. Nanometer scale mapping of the density of states in an inhomogeneous superconductor. *Europhys. Lett.* **54**, 84–90 (2001).
- Pan, S. H. *et al.* Microscopic electronic inhomogeneity in the high- T_c superconductor $\text{Bi}_2\text{Sr}_2\text{CaCu}_2\text{O}_{8+x}$. *Nature* **413**, 282–285 (2001).
- Ovchinnikov, Y. N., Wolf, S. A. & Kresin, V. Z. Intrinsic inhomogeneities in superconductors and the pseudogap phenomenon. *Phys. Rev. B* **63**, 064524–1–064524–6 (2001).

- Ghosal, A., Randeria, M. & Trivedi, N. Inhomogeneous pairing in highly disordered s-wave superconductors. *Phys. Rev. B* **65**, 014501–1–014501–13 (2002).
- Burgy, J., Mayr, M., Martin-Mayor, V., Moreo, A. & Dagotta, E. Colossal effects in transition metal oxides caused by intrinsic inhomogeneities. *Phys. Rev. Lett.* **87**, 277202–1–277202–4 (2001).
- Wang, Z., Engelbrecht, J. R., Wang, S., Ding, H. & Pan, S. H. Inhomogeneous d-wave superconducting state of a doped Mott insulator. Preprint cond-mat/0107004 at (<http://xxx.lanl.gov>) (2001); *Phys. Rev. B* **65**, (2002) (in the press).
- Martin, I. & Balatsky, A. V. Doping-induced inhomogeneity in high- T_c superconductors. *Physica C* **357–360**, 46–48 (2001).
- Wang, Q.-H., Han, J. H. & Lee, D.-H. Pairing near the Mott insulating limit. *Phys. Rev. B* **65**, 054501–1–054501–4 (2002).
- Phillips, J. C. & Jung, J. Nanodomain structure and function of high-temperature superconductors. *Phil. Mag. B* **81**, 745–756 (2001).
- Deutscher, G. & Müller, K. A. Origin of superconductive glassy state and extrinsic critical currents in high- T_c oxides. *Phys. Rev. Lett.* **59**, 1745–1747 (1987).
- Corson, J., Orenstein, J., Oh, S., O'Donnell, J. & Eckstein, J. N. Nodal quasiparticle lifetime in the superconducting state of $\text{Bi}_2\text{Sr}_2\text{CaCu}_2\text{O}_{8+x}$. *Phys. Rev. Lett.* **85**, 2569–2572 (2000).
- Barabash, S., Stroud, D. & Hwang, I.-J. Conductivity due to classical phase fluctuations in a model for high- T_c superconductors. *Phys. Rev. B* **61**, R14924–R14927 (2000).
- Loram, J. W., Luo, J. L., Cooper, J. R., Liang, W. Y. & Tallon, J. L. The condensation energy and pseudogap energy scale of $\text{Bi}_2\text{2212}$ from the electronic specific heat. *Physica C* **341–348**, 831–834 (2000).
- Takigawa, M. & Mitzi, D. B. NMR studies of spin excitations in superconducting $\text{Bi}_2\text{Sr}_2\text{CaCu}_2\text{O}_{8+x}$ single crystals. *Phys. Rev. Lett.* **73**, 1287–1290 (1994).
- Fong, H. F. *et al.* Neutron scattering from magnetic excitations in $\text{Bi}_2\text{Sr}_2\text{CaCu}_2\text{O}_{8+x}$. *Nature* **398**, 588–591 (1999).
- Vobornik, I. *et al.* Alternative pseudogap scenario: Spectroscopic analogies between underdoped and disordered $\text{Bi}_2\text{Sr}_2\text{CaCu}_2\text{O}_{8+x}$. *Phys. Rev. B* **61**, 11248–11250 (2000).
- Krasnov, V. M., Kovalev, A. E., Yurgens, A. & Winkler, D. Magnetic field dependence of the superconducting gap and the pseudogap in $\text{Bi}_2\text{2212}$ and $\text{HgBr}_2\text{-Bi}_2\text{2212}$, studied by intrinsic tunneling spectroscopy. *Phys. Rev. Lett.* **86**, 2657–2660 (2001).
- Suzuki, M. & Watanabe, T. Discriminating the superconducting gap from the pseudogap in $\text{Bi}_2\text{Sr}_2\text{CaCu}_2\text{O}_{8+x}$ by interlayer tunneling spectroscopy. *Phys. Rev. Lett.* **85**, 4787–4790 (2000).

Acknowledgements

We thank A. V. Balatsky, E. Dagotto, M. E. Flatté, S. A. Kivelson, V. Z. Kresin, R. B. Laughlin, J. W. Loram, D.-H. Lee, P. A. Lee, I. Martin, D. K. Morr, S. H. Pan, D. Pines, D. J. Scalapino, Z.-X. Shen, N. Trivedi and S. A. Wolf for discussions and communications. This work was supported by the LDRD program of Lawrence Berkeley National Laboratory, by the ONR, by the CULAR program of Los Alamos National Laboratory, by the Miller Research Foundation (J.C.D.), by IBM (K.M.L.), by Grant-in-Aid for Scientific Research, by a COE grant from the Ministry of Education, and by an International Joint Research Grant from NEDO (Japan).

Correspondence and requests for materials should be addressed to J.C.D. (e-mail: jcdavis@socrates.berkeley.edu).

Nitrogen loss from unpolluted South American forests mainly via dissolved organic compounds

Steven S. Perakis*† & Lars O. Hedin*†

* Department of Ecology and Evolutionary Biology, Corson Hall, Cornell University, Ithaca, New York 14853, USA

Conceptual^{1–4} and numerical^{5–8} models of nitrogen cycling in temperate forests assume that nitrogen is lost from these ecosystems predominantly by way of inorganic forms, such as nitrate and ammonium ions. Of these, nitrate is thought to be particularly mobile, being responsible for nitrogen loss to deep soil and stream waters. But human activities—such as fossil fuel combustion, fertilizer production and land-use change—have substantially altered the nitrogen cycle over large regions⁹, making it difficult to separate natural aspects of nitrogen cycling from those

† Present addresses: US Geological Survey, Forest and Rangeland Ecosystem Science Center, 3200 SW Jefferson Way, Corvallis, Oregon 97331, USA (S.S.P.); Department of Ecology and Evolutionary Biology, and Princeton Environmental Institute, Guyot Hall, Princeton University, Princeton, New Jersey 08544, USA (L.O.H.).

induced by human perturbations¹⁰. Here we report stream chemistry data from 100 unpolluted primary forests in temperate South America. Although the sites exhibit a broad range of environmental factors that influence ecosystem nutrient cycles^{11–13} (such as climate, parent material, time of ecosystem development, topography and biotic diversity), we observed a remarkably consistent pattern of nitrogen loss across all forests. In contrast to findings from forests in polluted regions, streamwater nitrate concentrations are exceedingly low, such that nitrate to ammonium ratios were less than unity, and dissolved organic nitrogen is responsible for the majority of nitrogen losses from these forests. We therefore suggest that organic nitrogen losses should be considered in models of forest nutrient cycling, which could help to explain observations of nutrient limitation in temperate forest ecosystems.

Hydrologic nutrient losses can constrain the accumulation and availability of nitrogen in terrestrial ecosystems¹⁰, with important long-term effects on productivity and carbon storage in nitrogen-limited temperate forest ecosystems^{5–8,14,15}. However, much of our understanding of nitrogen cycles in temperate forests has been developed from studies of relatively polluted and/or disturbed areas of the Northern Hemisphere. We here seek to characterize the limits to natural variations in nitrogen losses from forested ecosystems that have experienced minimal nitrogen pollution and minimal disturbances from logging and other human activities. We sampled hydrologic nitrogen losses from small watersheds that differed markedly in the principal state factors^{11–13} thought to influence how ecosystem nutrient cycles develop over time: climate, geologic parent material, time, topography and biota. We specifically considered whether broad variations in these state factors could yield appreciable variation in pathways of nitrogen loss for primary (that is, never logged) temperate forests that have developed across a large, yet unpolluted geographical region (Fig. 1).

We sampled 100 first-order streams that drained small

watersheds^{1,10} within 13 different geographical areas ($n = 2–14$ streams per area) between latitudes 40° S and 54° S in Chile and Argentina (Fig. 1). This region receives little or no nitrogen input from anthropogenic sources¹⁶, and encompasses wide variations in state factors that produce an array of temperate forest ecosystems^{11,17–21} (Table 1). Climate ranged between 4 and 11 °C mean annual temperature, and 500 to 5,840 mm mean annual precipitation across the different study areas^{18,19}. Geologic parent material varied widely, from highly metamorphosed schists in the Coastal Cordillera of Chile (study areas HU, PE, CP, MF), to glacial till in low-lying southern coastal regions (SC, TF), to recently deposited tephra over glacial till (CE, LP) or igneous rocks (SK, LF) in the Argentinean Andes, to deep tephra in the northern Lake District Chilean Andes (AN, PU), and a mixture of basalts, andesites and granites in the southern Lake District Chilean Andes (AA)²⁰. Time of ecosystem development ranged from $\leq 4,000$ years since retreat of glaciers at TF and volcanic activity at AN, to $\leq 18,000$ years at PY, SC, AA, CE, SK, LF and LP since retreat of the Llanquihue glacier, to $\gg 20,500$ years in the unglaciated areas HU, PE, CP and MF²⁰. Old-growth forests with little or no evidence of human disturbance dominated the watersheds in all but one area; LF watersheds contained mixtures of young and old forests recovering from natural and human-ignited fires. Areas were dominated by either one (TF, CE) or two (LP, SK) species of deciduous broadleaf tree, a single deciduous and single evergreen broadleaf species (AN), two species of coniferous evergreen trees (CP), or by higher diversity (greater than 3 species) mixtures of broadleaf evergreen (HU) or mixed coniferous and broadleaf evergreen (MF, PE, AA, PU, SC, LF) trees²¹.

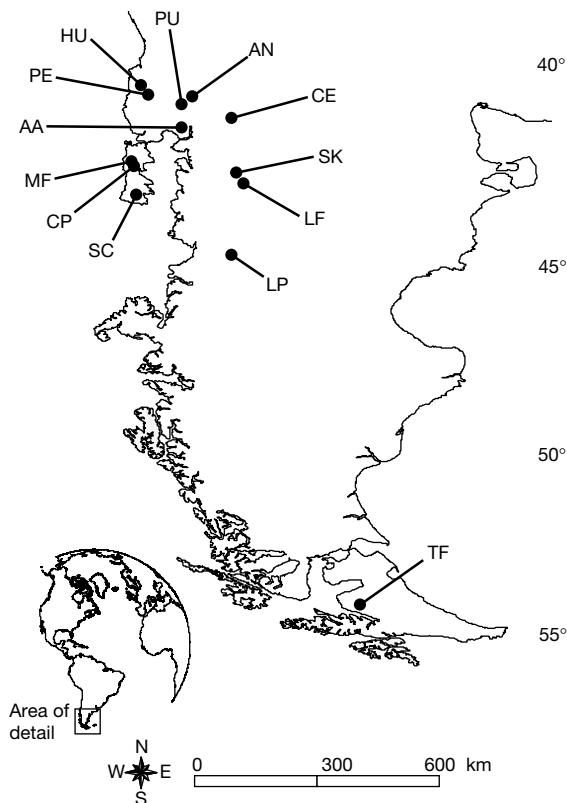


Figure 1 Map showing locations of study areas in South America.

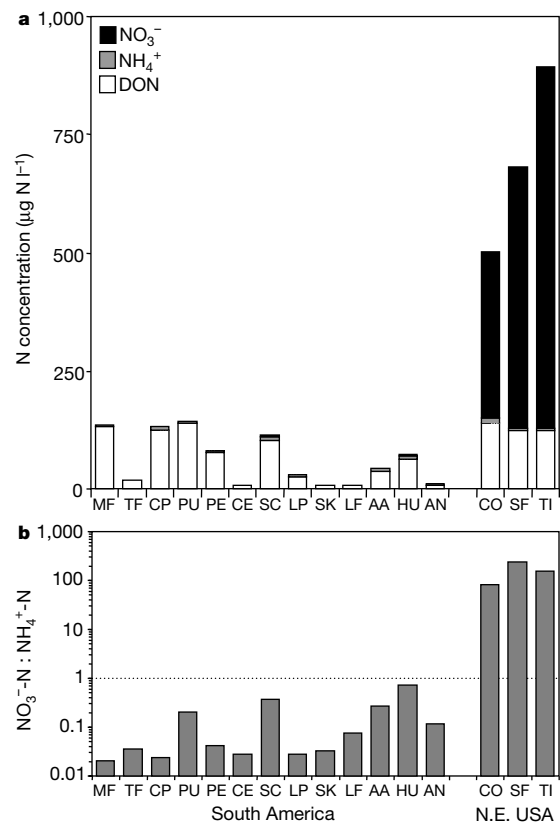


Figure 2 Hydrologic nitrogen losses from temperate forest watersheds in 13 areas of South America and 3 areas of eastern North America. **a**, Average concentrations of NH_4^+ , NO_3^- and DON; **b**, Ratios of NO_3^- -N : NH_4^+ -N. Sites arranged from left to right in order of increasing contribution of NO_3^- -N to total dissolved nitrogen. Dashed line in **b** indicates equal concentrations of NO_3^- -N and NH_4^+ -N.

We also sampled streams draining five old-growth forests from three chronically polluted areas of eastern North America. Watersheds located in the Tionesta National Forest in north-central Pennsylvania (TI) contained *Tsuga-Fagus* forests developed on sandstone²². Watersheds in Great Smoky Mountains National Park in eastern Tennessee²³ contained *Liriodendron-Fagus* (CO) and *Picea-Abies* forests (SF) developed on sandstone. All five North American watersheds escaped the Last Glacial Maximum.

We found that nitrate (NO_3^-) was consistently only a small fraction (average 5%, range 0.1–18%) of total dissolved nitrogen losses from unpolluted forests in southern Chile and Argentina, comprising less than 5% in 10 of 13 areas (Fig. 2a). Nitrate concentrations were uniformly low in the 13 different sample areas, ranging from 0.02 to $7.7 \mu\text{g N l}^{-1}$ (average $1.9 \mu\text{g N l}^{-1}$). Concentrations of dissolved ammonium (NH_4^+) were also low (average $4.9 \mu\text{g N l}^{-1}$, range 0.5–11 $\mu\text{g N l}^{-1}$) and contributed 3–36% (average 15%) of total nitrogen. In contrast to these inorganic forms of nitrogen, dissolved organic nitrogen (DON) was the dominant vector of hydrologic nitrogen loss in all sampled areas of southern Chile and Argentina, accounting for 61–97% of total N (average 80%). The range in DON concentrations among areas (8–135 $\mu\text{g N l}^{-1}$) was much greater than for dissolved inorganic nitrogen (DIN; range 0.5–18 $\mu\text{g N l}^{-1}$), with lowest DON ($< 26 \mu\text{g N l}^{-1}$) restricted to dry (mean annual precipitation $< 1,500$ mm; LP, SK, CE, LF) or young ($< 4,000$ years; TF and AN) areas.

The overall pattern of nitrogen loss from these South American forests contrasts sharply with losses from chronically polluted old-growth forests in eastern North America (Fig. 2a). Nitrate dominated (70–86%) total dissolved nitrogen losses from all North American watersheds, with concentrations (350–766 $\mu\text{g N l}^{-1}$) that were between 40 and 95 times greater than the highest concentrations among the South American forest areas. In addition, the average ratio of nitrate to ammonium expressed as N ($\text{NO}_3^- \text{N} : \text{NH}_4^+ \text{N}$) was very low (< 1) in all Chilean and Argentinean forests compared to ratios in the primary North American forests considered here (> 80 ; Fig. 2b), and compared to ratios in primary and secondary forests

elsewhere in eastern North America^{23–26}. Concentrations of DON did not differ significantly between unglaciated areas of South and North America ($P = 0.17$, t -test), although DON comprised a much smaller proportion of total nitrogen (20%) in North American forests. Variations in DON across South American watersheds were positively related to dissolved organic carbon (DOC; $r^2 = 0.70$, $P < 0.001$), and from the linear regression slope we estimate a DOC:DON ratio of 61 (s.e. = 4.0, $n = 100$) in hydrologic losses (see Supplementary Information). In comparison, DOC:DON ratios in losses from old-growth eastern North American forests are substantially lower (median 27, $n = 9$ watersheds^{24,25}), possibly as a result of chronic nitrogen deposition²⁶.

The pattern of low NO_3^- , high DON losses across South American forests is further supported by analyses of > 30 watershed streams¹⁰, and by five years of continuous measures of hydrology, stream nutrient losses (sampled twice a month), soil N transformations²⁷, and 24 soil lysimeters within four watersheds at the MF and CP sites (Table 1). These intensive local measures have shown that variations in relative losses of NO_3^- , NH_4^+ and DON across seasons and along soil-to-stream flow paths are small (less than about 25%) when compared to differences between South and North American forests (Fig. 2).

Our results show that narrow and characteristic patterns of nitrogen loss ($\text{DON} \gg \text{NH}_4^+ \text{N} > \text{NO}_3^- \text{N}$) can emerge throughout large geographical regions of unpolluted temperate forests, despite considerable variation in natural ecosystem state factors. These loss patterns differ substantially from traditional inorganic-based models and support studies that have previously identified DON as a vector of N loss from forests^{10,24–26,28}. Our findings have implications for understanding nitrogen fluxes at scales of ecosystems to landscapes in regions with minimal human impact. Using average N concentrations ($65 \mu\text{g N l}^{-1}$; range 11–140 $\mu\text{g N l}^{-1}$) from our sample areas, and assuming that evapotranspiration scales linearly with annual precipitation, we conservatively estimate that N losses range from 0.2 to $3.5 \text{ kg N ha}^{-1} \text{ yr}^{-1}$ (minimum 0.03 and maximum $7.5 \text{ kg N ha}^{-1} \text{ yr}^{-1}$) as a function of precipitation amount

Table 1 Characteristics of study areas

Study area	Code	Watersheds sampled	Longitude Latitude	Elevation (m)	MAT* (°C)	MAP† (mm)	Dominant vegetation	Age‡ (kyr)	Parent material§
Piuchué mixed forest	MF	10	42° 21' S 74° 06' W	500	6	5,840	<i>Nothofagus nitida</i> , <i>Drimys winteri</i> , <i>Podocarpus nubigena</i> , <i>Laurelia philippiana</i> , <i>Saxo-Gothaea conspicua</i>	$\gg 20.5$	PC
Cordillera Piuchué	CP	2	42° 22' S 74° 03' W	600	6	5,840	<i>Fitzroya cupressoides</i> , <i>Pilgerodendron uviferum</i>	$\gg 20.5$	PC
Tierra del Fuego	TF	2	54° 20' S 69° 15' W	100	4	500	<i>Nothofagus pumilio</i>	< 4	GL
Cordillera Pelada	PE	13	40° 10' S 73° 35' W	800	8	4,000	<i>N. nitida</i> , <i>D. winteri</i> , <i>P. nubigena</i> , <i>S. conspicua</i> , <i>F. cupressoides</i>	$\gg 20.5$	PC
Puyehue	PU	11	40° 46' S 72° 17' W	600	9	3,600	<i>P. nubigena</i> , <i>S. conspicua</i> , <i>D. winteri</i> var. <i>andina</i> , <i>Nothofagus dombeyi</i>	< 18	T
Southern Chiloé	SC	5	42° 52' S 73° 49' W	200	8	2,500	<i>N. nitida</i> , <i>S. conspicua</i> , <i>L. philippiana</i> , <i>P. nubigena</i>	< 18	GL
Antillanca	AN	6	40° 47' S 72° 11' W	1,100	4	5,400	<i>N. pumilio</i> , <i>Nothofagus betuloides</i>	< 4	T
Alerce Andino	AA	8	41° 32' S 72° 30' W	600	7	4,000	<i>N. nitida</i> , <i>Luma apiculata</i> , <i>S. conspicua</i> , <i>L. philippiana</i> , <i>D. winteri</i> var. <i>andina</i>	< 18	I
Hueicolla	HU	7	40° 09' S 73° 37' W	300	11	2,500	<i>N. nitida</i> , <i>D. winteri</i> , <i>L. apiculata</i> , <i>Amomyrtus luma</i> , <i>Weinmannia trichosperma</i>	$\gg 20.5$	PC
Lago La Plata	LP	9	44° 50' S 71° 44' W	1,000	7	770	<i>N. pumilio</i> , <i>Nothofagus antarctica</i>	< 18	T, GL
Lago Futralquén	LF	14	42° 51' S 71° 38' W	600	9	1,200	<i>Austrocedrus chilensis</i> , <i>N. dombeyi</i> , <i>Lomatia hirsuta</i>	< 18	T, I
Senda Kruger	SK	5	42° 50' S 71° 41' W	1,000	7	1,200	<i>N. pumilio</i> , <i>N. antarctica</i>	< 18	T, I
Lago Los Cesares	CE	8	41° 18' S 71° 41' W	1,150	7	1,500	<i>N. pumilio</i>	< 18	T, GL

* MAT, mean annual temperature.

† MAP, mean annual precipitation.

‡ Substrate age in kyr. $\gg 20.5$ indicates that the area escaped the Llanquihue glacial (18–20.5 kyr before present).

§ PC, pre-Cambrian mica-schists; GL, glacial till; T, tephra; I, igneous basalt, granite and andesite.

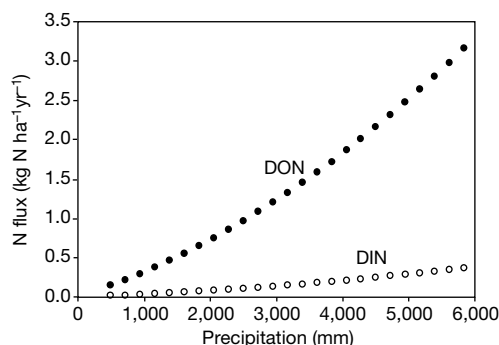


Figure 3 Hydrologic losses of dissolved inorganic and organic nitrogen, DIN and DON, estimated across the range of precipitation inputs to unpolluted forests of South America. Rates of hydrologic N flux from forest ecosystems were calculated by multiplying concentrations of DIN (average $6.8 \mu\text{g N l}^{-1}$) and DON (average $58.6 \mu\text{g N l}^{-1}$) by estimates of water flux. We calculated water fluxes as precipitation volume minus potential evapotranspiration losses. We assumed that percentage losses due to evapotranspiration decreased linearly from 50% to 10% across the range of precipitation from 500 to 5,840 mm.

(500–5,840 mm). We show in Fig. 3 that these rates of nitrogen loss are controlled by export of organic nitrogen compounds (DON), with little contribution from dissolved inorganic compounds (DIN). Our intensive hydrologic monitoring of watersheds at CP and MF further support these patterns; > 90% of nitrogen losses occurred as DON, and volume-weighted losses of nitrogen averaged 5.7 (s.e. = 0.09 , $n = 3$ years) and 5.9 (s.e. = 0.50 , $n = 3$ years) $\text{kg ha}^{-1} \text{yr}^{-1}$, respectively.

We conclude that unpolluted forests can display unexpectedly high rates of N output, primarily in the form of DON. We expect the influence of DON loss on ecosystem N balances to increase in importance along gradients from dry to moist forests (Fig. 3). Over longer periods, such nitrogen losses must be balanced by inputs from either nitrogen fixation²⁹ or natural atmospheric deposition³⁰. Our results suggest that DON losses from unpolluted forests can rival rates of NO_3^- -N loss in regions^{1–4,23–25} where N balances have been strongly biased in favour of recent anthropogenic inputs. Our findings support the idea^{10,31} that in regions with naturally low inputs of nitrogen from atmospheric deposition and nitrogen fixation, export of DON offers a mechanism to explain why nitrogen commonly limits plant productivity and carbon sequestration in moist temperate forests. We consider that models of forest ecosystems therefore should include DON as a central pathway of nitrogen loss, and should take into account the effects that such losses may have on nutrient limitation and carbon balances^{5–8}. □

Methods

We intensively sampled the chemistry and hydrology of small watershed streams at CP and MF every other week for over 5 years (1993–98). For the broader comparison between areas reported here, we included only results from December 1994 and January 1995, to coincide with collection efforts throughout southern Chile. However, results from these periods do not differ from long-term averages at CP and MF. For other locations in Chile, we collected duplicate samples from small watershed streams in December 1994 and January 1995, or in March 1997 for sites in the TF area. We sampled locations in Argentina in January and February 2001. Samples were filtered in the field through rinsed Gelman A/E glass-fibre filters ($< 1\text{-}\mu\text{m}$ nominal pore size, ref. 10) into 60-ml clean and leached polyethylene bottles. One replicate was immediately treated with 0.2 ml CHCl_3 to prevent biological activity, and all samples were kept cool and dark until analysis at Cornell University, generally within 2 weeks of collection. TI samples are averages from collections on October 1994 and May 1995. Data from CO and SF are averages of 9–23 sample collections per stream²³ throughout the entire year of 1998. These samples were stored frozen, and subsequently thawed and filtered through Gelman A/E glass-fibre filters. Details of analytical methods are given in ref. 10: NH_4^+ was analysed by Alpkem continuous-flow colorimetry, NO_3^- by Dionex ion chromatography, total dissolved nitrogen by colorimetry as NO_3^- following high-temperature persulphate digestion, and DOC by Shimadzu high-temperature platinum combustion. DON was calculated as total dissolved nitrogen minus NH_4^+ -N and NO_3^- -N.

Received 21 August; accepted 5 December 2001.

- Likens, G. E. & Bormann, F. H. *Biogeochemistry of a Forested Ecosystem* 2nd edn (Springer, New York, 1995).
- Aber, J. *et al.* Nitrogen saturation in temperate forest ecosystems: Hypotheses revisited. *BioScience* **48**, 921–34 (1998).
- Tamm, C. O. *Nitrogen in Terrestrial Ecosystems* (Springer, Berlin, 1991).
- Stoddard, J. L. in *Environmental Chemistry of Lakes and Reservoirs* (ed. Baker, L. A.) 223–284 (American Chemical Society, Washington DC, 1994).
- Schimel, D. S., Braswell, B. H. & Parton, W. J. Equilibration of the terrestrial water, nitrogen, and carbon cycles. *Proc. Natl Acad. Sci. USA* **94**, 8280–8283 (1997).
- Rastetter, E. B. *et al.* Resource optimization and symbiotic nitrogen fixation. *Ecosystems* **4**, 369–388 (2001).
- McGuire, A. D. *et al.* Equilibrium responses of global net primary production and carbon storage to doubled atmospheric carbon dioxide: sensitivity to changes in vegetation nitrogen concentration. *Glob. Biogeochem. Cycles* **11**, 173–189 (1997).
- McKane, R. B. *et al.* Climatic effects on tundra carbon storage inferred from experimental data and a model. *Ecology* **78**, 1170–1187 (1997).
- Vitousek, P. M. *et al.* Human alteration of the global nitrogen cycle: Sources and consequences. *Ecol. Appl.* **7**, 737–750 (1997).
- Hedin, L. O., Armesto, J. J. & Johnson, A. H. Patterns of nutrient loss from unpolluted, old-growth temperate forests: Evaluation of biogeochemical theory. *Ecology* **76**, 493–509 (1995).
- Jenny, H. *Factors of Soil Formation* (McGraw-Hill, New York, 1941).
- Gorham, E. Factors influencing supply of major ions to inland waters, with special reference to the atmosphere. *Geol. Soc. Am. Bull.* **72**, 795–840 (1961).
- Vitousek, P. M. & Reiners, W. A. Ecosystem succession and nutrient retention: a hypothesis. *BioScience* **25**, 376–381 (1975).
- Nadelhoffer, K. J. *et al.* Nitrogen deposition makes a minor contribution to carbon sequestration in temperate forests. *Nature* **398**, 145–148 (1999).
- Oren, R. *et al.* Soil fertility limits carbon sequestration by forest ecosystems in a CO_2 -enriched atmosphere. *Nature* **411**, 469–472 (2001).
- Holland, E. A., Dentener, F. J., Braswell, B. H. & Sulzman, J. M. Contemporary and pre-industrial global reactive nitrogen budgets. *Biogeochemistry* **46**, 7–43 (1999).
- Alaback, P. Comparative ecology of temperate rainforests of the Americas along analogous climatic gradients. *Rev. Chil. Hist. Nat.* **64**, 399–412 (1991).
- Almeyda, A. E. & Saez, S. F. *Recopilación de Datos Climáticos de Chile y Mapas Sinópticos Respetivos* (Ministerio de Agricultura, Santiago, Chile, 1958).
- Paruelo, J. M., Beltrán, A., Jobbagy, E. G., Sala, O. E. & Golluscio, R. A. The climate of Patagonia: General patterns and controls on biotic processes. *Ecologia Austral* **8**, 85–101 (1998).
- Villagrán, C. (ed.) *The Quaternary of the Lake District of Southern Chile*. International Workshop “The Quaternary of Chile” (IGCP-281, Latin American Botanical Network, Santiago, Chile, 1993).
- Veblen, T. T., Donoso, C., Kitzberger, T. & Rebertus, A. J. in *The Ecology and Biogeography of Nothofagus Forests* (eds Veblen, T. T., Hill, R. S. & Read, J.) 293–353 (Yale Univ. Press, New Haven, 1996).
- Aguilar, R. & Arnold, R. W. Soil-landscape relationships of a climax forest in the Allegheny High Plateau, Pennsylvania. *Soil Sci. Soc. Am. J.* **49**, 695–701 (1985).
- Nodvin, S. C., Van Miegroet, H., Lindberg, S. E., Nicholas, N. S. & Johnson, D. W. Acidic deposition, ecosystem processes, and nitrogen saturation in a high elevation southern Appalachian watershed. *Wat. Air Soil Pollut.* **85**, 1647–1652 (1995).
- Goodale, C. L., Aber, J. D. & McDowell, W. H. The long-term effects of disturbance on organic and inorganic nitrogen export in the White Mountains, New Hampshire. *Ecosystems* **3**, 433–450 (2001).
- Campbell, J. L. *et al.* Dissolved organic nitrogen budgets for upland, forested ecosystems in New England. *Biogeochemistry* **49**, 123–142 (2000).
- McDowell, W. H., Currie, W. S., Aber, J. D. & Yano, Y. Effects of chronic nitrogen amendments on production of dissolved organic carbon and nitrogen in forest soils. *Wat. Air Soil Pollut.* **105**, 175–182 (1998).
- Perakis, S. S. & Hedin, L. O. Fluxes and fates of nitrogen in soil of an unpolluted old-growth temperate forest, southern Chile. *Ecology* **82**, 2245–2260 (2001).
- Sollins, P. *et al.* The internal element cycles of an old-growth Douglas-fir ecosystem in western Oregon. *Ecol. Monogr.* **50**, 261–285 (1980).
- Cleveland, C. C. *et al.* Global patterns of terrestrial biological nitrogen (N_2) fixation in natural ecosystems. *Glob. Biogeochem. Cycles* **13**, 623–645 (1999).
- Weathers, K. C., Lovett, G. M., Likens, G. E. & Caraco, N. F. M. Cloudwater inputs of nitrogen to forest ecosystems in southern Chile: Forms, fluxes, and sources. *Ecosystems* **3**, 590–595 (2000).
- Vitousek, P. M., Hedin, L. O., Matson, P. A., Fownes, J. H. & Neff, J. in *Successes, Limitations, and Frontiers in Ecosystem Science* (eds Pace, M. L. & Groffman, P. M.) 432–451 (Springer, New York, 1998).

Supplementary Information accompanies the paper on Nature's website (<http://www.nature.com>).

Acknowledgements

We thank S. Nodvin for sharing samples from CO and SF; M. Medina, B. Houlton, J. Armesto, C. Perez, G. Steinhart, J. O'Brien, G. Likens, J. Franklin, O. Sala, G. Lewis, T. Plum, J. Jamison, Corporación Nacional Forestal de Chile and Administración de Parques Nacionales de Argentina for site selection, field assistance and sample collection; M. Brown and K. Lohse for sample analysis; E. Boyer for cartography; and J. Compton, T. Fahey, R. Howarth, K. Lajtha, J. von Fischer, G. Lewis and P. Vitousek for comments on the manuscript. This work is a contribution to the Cordillera Piuichú Ecosystem Study and was supported by the Andrew W. Mellon Foundation, the NASA Earth Systems Science Fellowship Program and the US NSF.

Correspondence and requests for materials should be addressed to S.S.P. (e-mail: sparakis@usgs.gov).

Estimating the human health risk from possible BSE infection of the British sheep flock

N. M. Ferguson, A. C. Ghani, C. A. Donnelly, T. J. Hagenaars & R. M. Anderson

Department of Infectious Disease Epidemiology, Faculty of Medicine, Imperial College of Science, Technology and Medicine, St Mary's Campus, Norfolk Place, London W2 1PG, UK

Following the controversial failure of a recent study¹ and the small numbers of animals yet screened for infection², it remains uncertain whether bovine spongiform encephalopathy (BSE) was transmitted to sheep in the past via feed supplements and whether it is still present. Well grounded mathematical and statistical models are therefore essential to integrate the limited and disparate data, to explore uncertainty, and to define data-collection priorities. We analysed the implications of different scenarios of BSE spread in sheep for relative human exposure levels and variant Creutzfeldt-Jakob disease (vCJD) incidence. Here we show that, if BSE entered the sheep population and a degree of transmission occurred, then ongoing public health risks from ovine BSE are likely to be greater than those from cattle, but that any such risk could be reduced by up to 90% through additional restrictions on sheep products entering the food supply. Extending the analysis to consider absolute risk, we estimate the 95% confidence interval for future vCJD mortality to be 50 to 50,000 human deaths considering exposure to bovine BSE alone, with the upper bound increas-

ing to 150,000 once we include exposure from the worst-case ovine BSE scenario examined.

The aim of this study was not to evaluate the probability that BSE has entered the sheep flock, but rather, given the pessimistic assumption that infection has occurred, to explore its potential extent and pattern of spread. In this, we used epidemiological parameter estimates from experimental BSE infections of sheep, and, where data are unavailable, assumed (given the observed similarities in BSE and scrapie pathogenesis in sheep) that other aspects of disease epidemiology resemble those of scrapie. Analyses were constrained to be consistent with the failure to detect the BSE agent in a small sample of 180 brains² collected between 1996 and 2000 from sheep diagnosed with scrapie (giving an upper bound for BSE prevalence within apparently scrapie-affected sheep of 2%; Fig. 1a), and are also broadly consistent with an assessment of historical exposure of the ovine population to meat and bonemeal (MBM)³.

Key to our analysis are estimates of the infectivity in animal tissues during disease incubation. Data are limited for BSE in sheep acquired by oral challenge, but using new experimental results and published data from studies of both scrapie⁴⁻⁷ and sheep BSE^{8,9} pathogenesis, we constructed an infectiousness profile. This profile was based on temporal changes in the density of the agent in different tissues, weighted by the proportion of such tissue in the host's body (Fig. 1b). This profile differs from that of BSE in cattle^{10,11}, with a more rapid rise in overall infectivity early in the incubation period in a wide range of tissues (for example, spleen and lymph nodes). The sheep-human infectiousness profiles (Fig. 1b) adjust for tissue-specific usage in food¹² and the effect of the 1997 specified risk materials (SRM) ban in sheep.

The distribution of the BSE incubation period in sheep is not well characterized, but on the basis of the limited available data, we used

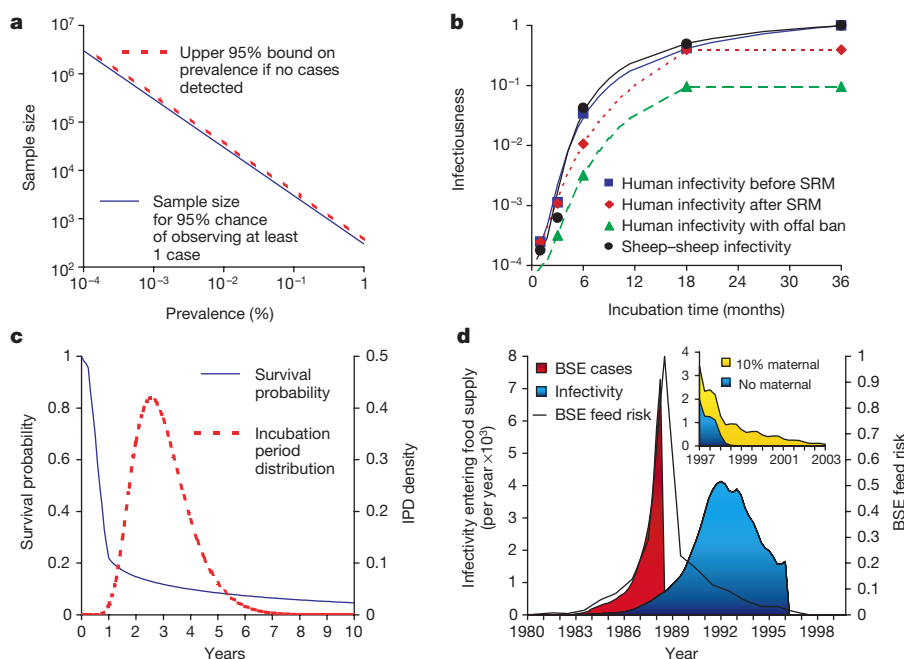


Figure 1 Epidemiological inputs to transmission model. **a**, Relationship between sample size and detectable prevalence in screening studies. **b**, Infectiousness of sheep as a function of time from infection (see Methods). Sheep and human profiles are separately normalized to give maxima of 1. **c**, Survival probability of sheep as a function of age, and assumed incubation period distribution (IPD) of BSE in sheep. The survivorship function was estimated from annual data from the June census, slaughter, export and disappearance statistics, and data on the seasonality of lamb slaughter. **d**, Before mid-1988, both reported and unreported clinical BSE cases could be used for food

(red-shaded curve). After that time, BSE was made notifiable and cases were destroyed, so we assume none entered food. The blue-shaded curve represents the rate of slaughter (per year) of pre-clinical infected cattle weighted by infectiousness relative to disease onset, assuming infectiousness grows at the exponential rate of 4 per year. The over-30-month scheme ended most bovine exposure in 1996, with estimates of residual levels (inset) being dependent on the extent of maternal transmission of BSE. The solid black curve represents the estimated infection hazard to cattle and sheep posed by infectivity in contaminated feed, relative to the maximum level reached in 1988.

an offset gamma distribution with a mean of 3 years (Fig. 1c) and substantial variance (intended to capture variation caused by dose dependency and host genotype) both for sheep infected by feed and those infected horizontally. Pathogenesis and susceptibility are dependent on the genetic background of the host^{13–15}, and genotype frequencies of the key polymorphisms vary considerably within and between flocks of different breeds^{16,17}. Collation of limited available data suggests that roughly one-third of sheep in Great Britain (England, Scotland and Wales) have BSE-susceptible genotypes (see Supplementary Information). Exposure of the sheep flock to the BSE agent via contaminated feed (Fig. 1d) is assumed to mirror (albeit at a much lower level) that of British cattle, as estimated in back-calculation analyses of the BSE epidemic^{18,19}. Exposure is likely to have occurred as far back as the early 1980s, before the disease was identified in cattle^{18,19}. As for BSE in cattle²⁰, host survivorship is important given the long incubation period of the disease. Best estimates (see Supplementary Information) are presented in Fig. 1c (giving a mean life expectancy of 1.5 years), although more precise data are urgently required, perhaps based on a sheep equivalent of the British Cattle Tracing System (<http://www.bcms.gov.uk>).

Capturing the information in Fig. 1 requires a framework that integrates the temporal evolution of BSE pathogenesis in the individual host (incorporating age-dependent susceptibility and exposure) into a mathematical model of transmission within (including seeding and spread) and between flocks. This model builds on previous analyses of within- and between-flock transmission of scrapie^{21,22}, and consists of a set of nonlinear partial differential equations detailing the transmission dynamics of the agent and the demography of the sheep flock under time-dependent exposure to BSE-contaminated feed. The dynamics of disease transmission within the sheep population are determined by the magnitude of the respective basic reproduction numbers of the agent within a flock (R_0^A) and between flocks (R_0^F). The reproduction

numbers define the average number of secondary cases or flocks generated by one primary case or infected flock in a susceptible flock or population of flocks. We considered three representative scenarios: (I) $R_0^A > 1$, $R_0^F < 1$ —self-sustaining transmission within a flock but not between flocks; (II) $R_0^A > 1$, $R_0^F > 1$ —the worst-case scenario for future spread, with spread within and between flocks inducing an expanding epidemic; and (III) $R_0^A < 1$, $R_0^F < 1$ —the best-case scenario, with non-self-sustaining transmission both within and between flocks (see Fig. 2 for precise parameter values).

For each scenario, the level of flock infection due to MBM exposure was adjusted to give BSE prevalence below 2% of scrapie prevalence at present (consistent with the results of ongoing studies screening sheep brains). Judgements of scenario consistency therefore depend on the estimated prevalence of scrapie in British sheep, with such estimates being based on limited data from detailed surveys of specific flocks (using clinical criteria for diagnosis) and a postal survey of farmers intended to characterize national historical patterns of scrapie incidence^{23,24}. The uncertainties in interpreting these data (giving estimates of infection prevalence anywhere between 0.1% and 1%, see Supplementary Information) make large-scale (Fig. 1a) screening of the national flock for transmissible spongiform encephalopathies (TSEs) a priority. In constructing the scenarios, we assumed a scrapie prevalence of about 0.3%, with scenarios II and III then corresponding to a BSE prevalence of 0.5% that of scrapie, and scenario I to a prevalence of about 2% that of scrapie. Scenario I was thus intended to represent something of a worst case in terms of the numbers of animals infected to date, although we cannot exclude the possibility of even larger epidemics given the limited data currently available.

Figure 2 displays the epidemiological characteristics of these scenarios in terms of within-flock and overall prevalence (Fig. 2a–c), and their implications for human exposure via food (Fig. 2d–f). The estimates of exposure incorporate data on human

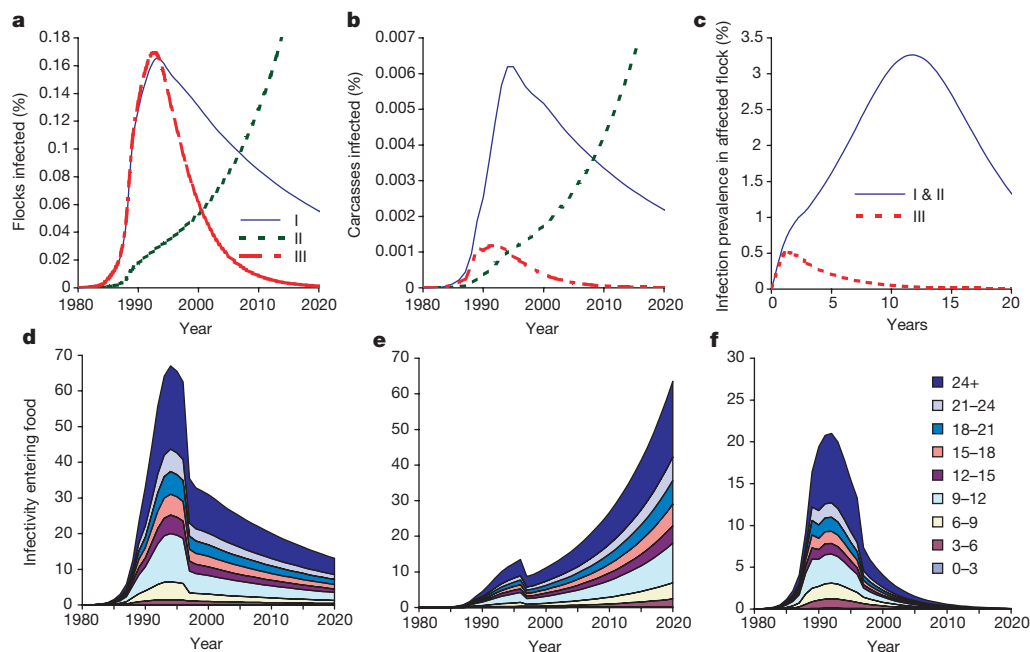


Figure 2 Epidemiological characteristics of BSE transmission scenarios in sheep. $R_0^A = 2$, $R_0^F = 0.8$ and $\beta_B = 0.2\%$ per year for scenario I; $R_0^A = 2$, $R_0^F = 1.5$ and $\beta_B = 0.025\%$ per year for scenario II; and $R_0^A = 0.8$, $R_0^F = 0.5$ and $\beta_B = 0.2\%$ per year for scenario III; where β_B is the assumed flock infection incidence rate per unit of feed risk profile shown in Fig. 1c. By comparison, β_B values of 0.2% and 0.025% represent per-animal infection hazards about 50- and 400-fold less than that experienced by cattle. **a**, Proportion of flocks affected through time for scenarios I–III. **b**, Proportion of carcasses entering the

food supply infected (at any incubation stage) with BSE. **c**, Prevalence in affected flock as a function of time since initial entry of infection into the flock. **d–f**, Estimated infectivity (in units of maximally infectious carcasses) entering the food supply under scenarios I, II and III, respectively, derived from the infectiousness profiles shown in Fig. 1a. Varying the proportion of flocks affected with BSE (shown in **a**, and determined by β_B) scales the exposure curves in **d–f** proportionately.

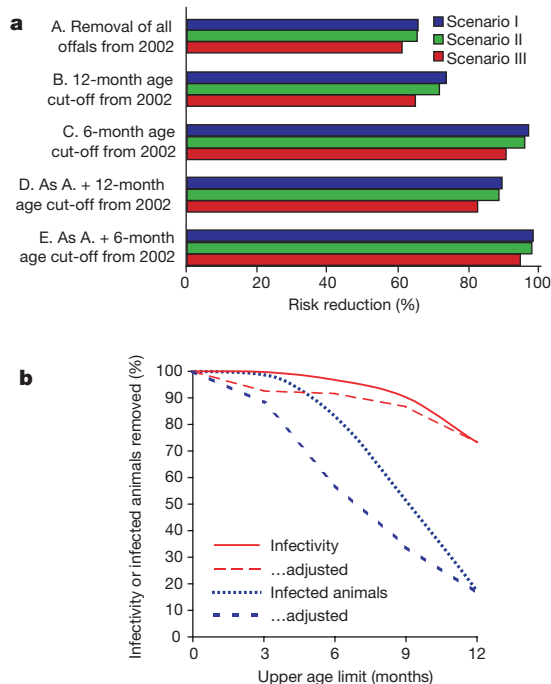


Figure 3 Impact of risk-reduction measures. **a**, Estimated relative impact of various measures on current exposure to BSE infectivity in sheep food products for scenarios I–III. ‘Removal of all offals’ corresponds to a ban on all internal organs and central nervous system tissue of sheep entering the human food supply, but does not assume removal of lymph nodes. Estimates presented were calculated assuming that age restrictions would not change slaughter patterns. **b**, Estimated impact of imposing age restrictions as a function of the upper age limit imposed for scenario I. Effects on infectivity and total number of infected animals entering the food supply are shown, with dashed curves showing how the impact is reduced if slaughter patterns are adjusted after such a measure, such that all animals currently slaughtered under 12 months old are then slaughtered under the upper age limit imposed.

consumption of ovine material (see Supplementary Information), which indicate that 67% of lambs and 83% of sheep older than 12 months slaughtered for consumption in 1999 were consumed domestically in the UK. Very few live sheep are imported into Great Britain, and most imported lamb meat originates from New Zealand, which has never detected signs of BSE or scrapie infection in either its bovine or ovine populations. Thus the potential risk from BSE-infected sheep arises from home-bred animals.

The sudden drop in human exposure in the late 1990s (Fig. 2d–f) is due to the SRM ban for sheep (banning the consumption of the

skull, tonsils and spinal column of animals older than 12 months and the spleen of all animals). In the worst-case scenario (II), the degree of risk is still rising steeply, whereas for the other scenarios risk is falling, approaching very low levels in the best case (III). Critically, even in the worst-case scenario, the great majority of past exposure of the human population to BSE arose from cattle (Fig. 1d). However, for all three scenarios considered (and therefore assuming BSE did enter the sheep flock), the number of maximally infectious sheep entering the food supply in Great Britain at present is estimated to be greater than the number of maximally infectious cattle^{18,19} (Fig. 1d, inset). This comparison equates the risk from one maximally infectious sheep with that of a cow, despite the difference in body mass, an assumption motivated by the evidence of more widespread distribution of infectivity in sheep.

Potential risk-reduction strategies include restrictions on the age of sheep slaughtered for consumption and enhanced tissue-based controls to reduce the amount of infectivity entering the food supply; additional measures based on flock history or ram genotype might also be possible but are not considered here. Combined tissue- and age-based restrictions are estimated to reduce current and future risk by at least 80% for all three sheep BSE scenarios considered (Fig. 3a). This is encouraging, but the precise values shown depend on the accuracy of the data on the development of infectivity in different tissues of BSE-infected sheep. Furthermore, verifying the age of sheep is difficult without an identification scheme, so in the short term it would be feasible only to exploit seasonal birthing patterns and dental indicators to impose approximate age restrictions. How the impact of 12-month age restrictions might be reduced by compensatory changes in slaughter patterns is shown in Fig. 3b.

Translation of the patterns of relative exposure through time into measures of absolute risk requires estimation of potential vCJD mortality in the human population of Great Britain. Given the uncertainty in the parameters determining such predictions (in particular, the incubation period distribution and human susceptibility), analyses must be based on the best available estimates of temporal changes in human exposure to infected material. The confidence bounds on vCJD mortality shown in Table 1 characterize human exposure by infectivity-weighted estimates of the numbers of BSE-infected cattle (Fig. 1d) and sheep (Fig. 2) slaughtered for consumption through time (correcting for early under-reporting of cattle BSE incidence). We modelled the vCJD epidemic solely in the 40% of the population that are methionine homozygous at prion protein (PrP) codon 129, assuming no other genetic variation in susceptibility. Currently no case data exist with which to constrain epidemic scenarios for other genotypes, but if future cases are diagnosed then upper bounds on epidemic size will increase. Compared with our previous estimates²⁵, upper bounds on epidemic size

Table 1 95% confidence intervals for future vCJD deaths

Cattle only*	2001–2002	2001–2005	2001–2010	2001–2020	2001–2040	2001–2080
A	20–100	40–400	40–1,200	40–5,000	40–20,000	50–50,000
B	20–100	40–400	40–1,200	40–5,000	40–20,000	40–40,000
C	20–100	30–400	40–1,400	40–7,000	40–40,000	40–90,000
D	20–100	30–400	30–1,300	40–7,000	40–40,000	40–100,000
E	20–100	40–350	40–1,100	40–5,000	40–20,000	50–35,000
Cattle and sheep†	Sheep:cattle infectivity ratio			2001–2020	2001–2040	2001–2080
Scenario I	1:1			40–5,000	40–20,000	40–60,000
	10:1			40–5,000	50–30,000	50–70,000
Scenario II	1:1			20–5,000	20–20,000	20–70,000
	10:1			40–5,000	70–30,000	110–150,000
Scenario III	1:1			40–5,000	40–20,000	40–50,000
	10:1			40–5,000	40–20,000	40–50,000

* Cattle-only calculations assume: A, BSE exposure in humans was proportional to the BSE cases before the mid-1988 in addition to infected animal slaughter rates shown in Fig. 1c; B, as A but excluding exposure to reported BSE cases; C, as A but additionally fitting to at least three vCJD deaths reported before 2001 being infected before 1986, as suggested by analysis of the Queenborough cluster; D, as A but assuming the 1989 specified bovine offal ban reduced human exposure by at least 80%; E, as A but restricting the mean incubation period to be no longer than 60 years.

† Calculations for cattle and sheep assume exposure as A from cattle plus exposure from sheep as in Fig. 2d–f from scenarios I–III, assuming sheep are equally or 10 times as infectious as cattle. Short-term predictions for these scenarios are similar to those excluding sheep.

in the absence of BSE in sheep have reduced (largely as a result of a change in statistical methods), being in the range 50,000–100,000 depending on the assumptions made. These values are substantially greater than recently published estimates derived assuming a cruder representation of past trends in human exposure to BSE infectivity^{26,27}. Best-fit estimates associated with 2001–2080 confidence bounds generally lie in the range 100–1,000, but the fits of the model vary little for fewer than 10,000 deaths. It should be noted (Table 1, E) that large epidemics are still possible even if the mean incubation period is less than 60 years. In the presence of BSE in sheep, the upper bound is substantially increased only if BSE is capable of becoming endemic in the national flock (scenario II).

Although the risk analysis presented here incorporates a wide variety of available information into a single integrated framework, its reliability depends on the quality and volume of data available for parameter estimation. The limited data highlight the need for further studies to measure: (1) scrapie and BSE prevalence in sheep (stratified by age), employing sample sizes sufficiently large to detect low prevalence; (2) sheep survivorship more precisely; (3) BSE infectivity in sheep quantitatively, by stage of incubation, tissue and sheep genotype; (4) age-dependent susceptibility to TSE infection in a variety of species, including sheep and cattle; and (5) historical trends in bovine and ovine tissue consumption. Molecular typing methods giving rapid results^{28,29} are clearly valuable for prevalence screening and strain typing (whether for BSE or scrapie), but study design should take account of test sensitivity and therefore consider which tissue should be tested (brain not necessarily being optimal). Given the uncertainty regarding the presence of BSE in the British sheep flock, such large-scale testing is a priority. In the interim, this analysis informs prevalence survey design and policy consideration of the potential benefits of additional risk-reduction measures. □

Methods

Additional detail is provided as Supplementary Information.

Modelling transmission of BSE in sheep

Infectiousness of a BSE-infected sheep as a function of time τ since infection (relative to the incubation period T) was estimated by fitting the parametric form $\rho(\tau/T) = \exp(a(\tau/T)^b)/((\tau/T)^b + c)$ to tissue-specific infectivity data from studies of pathogenesis of ovine BSE³⁰ and scrapie^{4–7}. The data were weighted by total tissue mass when estimating sheep-to-sheep infectiousness, and by the proportion of tissue mass entering food when estimating human exposure to infectivity (Fig. 1b). This approach captures the impact on overall infectiousness of between-tissue variation in PrP accumulation during pathogenesis of ovine BSE—namely, that some tissues (for example, lymph nodes) rapidly develop detectable infectivity, the growth of which later slows and/or saturates, whereas others (for example, brain) only exhibit high levels of infectivity close to clinical onset.

Because mass-action models cannot capture the observed clustering of cases²⁰, we developed a model of TSE transmission in sheep that incorporated three relevant tiers: (1) the individual animal—capturing age-dependent susceptibility and exposure, and pathogenesis of disease; (2) the individual flock—capturing within-flock sheep-to-sheep transmission; and (3) the national population of flocks—capturing between-flock transmission and exposure to contaminated feed.

Assuming homogenous mixing within a flock, the deterministic susceptible-infected model is:

$$\begin{aligned} \frac{dx}{dt} + \frac{\partial x}{\partial a} &= -(\Lambda(t)\kappa(a) + \mu(a))x & x(t, 0) &= B & \mu(a) &= -\frac{1}{S} \frac{dS}{da} \\ \frac{\partial y}{\partial t} + \frac{\partial y}{\partial a} + \frac{\partial y}{\partial \tau} &= -(F_A(\tau) + \mu(a))y & F_A(\tau) &= \frac{f_A(\tau)}{1 - \int_0^\tau f_A(\tau')d\tau'} \\ y(t, a, 0) &= \Lambda(t)\kappa(a)x(t, a) \\ \Lambda(t) &= \beta_A \int_{\tau=0}^t \int_{T=\tau}^\infty F_A(T)\rho(\tau/T)y(t, a, \tau)dT d\tau & y(0, a, 0) &= \frac{\kappa(a)Y_0}{\int \kappa(a)da} \end{aligned}$$

where we denote the densities of susceptible and infected animals of age a at time t by $x(t, a)$ and $y(t, a, \tau)$; time from infection by τ ; the incubation period distribution by $f_A(\tau)$ (see Fig. 1c); the force of infection by $\Lambda(t)$; the transmission coefficient for horizontal spread by β_A ; the fixed birth rate by B ; the mortality rate by $\mu(a)$; and age-dependent susceptibility by $\kappa(a)$ (conservatively assumed to be constant for the first 12 months of life and zero thereafter). The number of animals infected at the start of a flock outbreak is Y_0 .

Assuming that the infectiousness at time t of a flock infected at $t = 0$ to other flocks is proportional to the within-flock force of infection $\Lambda(t)$, we modelled transmission within a population of 100,000 flocks of 400 animals each. Genetic heterogeneity in susceptibility to infection was modelled at the flock level, with 33,000 flocks assumed to have all animals susceptible and the remainder completely resistant. The introduction of infection into a flock from an external source (contaminated feed or other sheep) was modelled as a rare event, initially infecting 1% of animals.

The infection dynamics of flocks were modelled using an SIRS framework, whereby flocks are initially 'susceptible' to infection, enter an 'infected' state of extended duration, 'recover' and become resistant to further infection (for a period of 20 years), then revert to the 'susceptible' state. The flock-level 'recovery' process approximates flock outbreak extinction mechanisms, such as demographic stochasticity (likely to be critical³⁰ given the low reported prevalence of scrapie), control measures and selection for scrapie-resistant genotypes.

Denoting the number of susceptible flocks at time t by $s(t)$, flocks infected time θ ago by $h(t, \theta)$ (with infectivity $\Lambda(\theta)$ from the within-flock model) and recovered/resistant flocks by $r(t)$, model dynamics are described by:

$$\begin{aligned} \frac{ds}{dt} &= \xi r - \Omega(t)s & \Omega(t) &= \beta_F \int_{\theta=0}^\infty \Lambda(\theta)h(t, \theta)d\theta + \beta_B\phi(t) \\ \frac{\partial h}{\partial t} + \frac{\partial h}{\partial \theta} &= -F_F(\theta)h & F_F(\theta) &= \frac{f_F(\theta)}{1 - \int_0^\theta f_F(\theta')d\theta'} \\ \frac{dr}{dt} &= \int F_F(\theta)h(t, \theta)d\theta - \xi r & s(0) &= N \quad h(t, 0) = \Omega(t)x \quad h(0, \theta) = 0 \end{aligned}$$

Here, $f_F(\theta)$ is the 'incubation period' distribution for flocks (gamma distributed with mean 6 years; see Supplementary Information); $\Omega(t)$ the force of infection for flocks at time t ; β_F and β_B the coefficients for between-flock transmission and exposure of flocks to contaminated feed; $\phi(t)$ the relative risk from contaminated feed at time t (estimates from refs 18, 19); and ξ ($= 0.05$ per year) the rate at which recovered flocks re-enter the susceptible pool. The values of β_A and β_F corresponding to required values of R_0^A and R_0^F were determined numerically.

Exposure of the human population to BSE infectivity in sheep was represented by the number of infected animals slaughtered for food (stratified by age and incubation stage, weighted by infectivity; Fig. 2d–f). The effectiveness of risk-reduction measures was evaluated by examining the distribution of exposure as a function of animal age, and by comparing results obtained from the infectivity profiles corresponding to current SRM controls and a ban on all offal.

It should be noted that this model framework reproduces observed within-flock and overall scrapie incidence patterns well if run to endemicity (results not shown).

Prediction of vCJD incidence

As in earlier work²⁵, the probability density that an individual develops clinical disease at time t and age a is

$$p(t, a) = S_H(t, a) \int_{t-a}^t f(t-u)I(u, a-t+u)\exp\left[-\int_0^u I(u', a-t+u')du'\right]du$$

where $S_H(t, a)$ is the probability that someone born at time $t-a$ will survive to age a (derived from census data) and $f(t-u)$ is the incubation period distribution (modified lambda distribution). The infection hazard is given by

$$I(t, a) = \beta g(a) \left(v_c(t) \int \Omega_c(z)\omega_c(z, t)dz + v_s(t) \int \Omega_s(z)\omega_s(z, t)dz \right)$$

where β is the transmission coefficient and $g(a)$ an age-dependent susceptibility/exposure function (uniform with gamma-distributed tails). For each infectious species i (C for cattle, S for sheep), $\Omega_i(z)$ is the relative infectiousness of an animal time z from disease onset, $v_i(t)$ is the time-dependent effectiveness of risk-reduction measures (for cattle, this is parameterized as a step reduction in 1989 due to the specified bovine offal ban), and $\omega_i(z, t)$ is the number of animals slaughtered for consumption stratified by time and incubation stage. (Values for sheep were obtained from the above model, and those for cattle used updated back-calculation estimates^{18,19}, with $\int \Omega_c(z)\omega_c(z, t)dz$ plotted in Fig. 1d.)

The relative infectiousness of cattle by incubation stage was assumed to increase exponentially from a baseline level to a maximum value before onset of clinical signs²⁵. We restricted analysis to the 40% of the population that are methionine homozygous at PrP codon 129, assuming no other genetic variation in susceptibility.

Through numerical solution of the inverse problem (see Supplementary Information for further details), β was calculated as a function of case incidence, allowing incidence of vCJD deaths in any time interval to be treated as a model parameter. This enabled nonlinear optimization techniques to be used to obtain likelihood profiles by fitting the model to the joint age- and time-stratified mortality data to the end of 2000. We obtained 95% confidence bounds from the one-dimensional likelihood profiles. Infection prevalence is poorly constrained by the observed incidence data, and can range from being equal to mortality to 100-fold larger, with little effect on model fit.

Received 21 November; accepted 12 December 2001.

Published online 9 January 2002, DOI 10.1038/nature709.

1. Frankish, H. Samples blunder renders sheep-BSE study useless. *Lancet* **358**, 1436 (2001).

2. Beckett, M. *House of Commons Hansard Debates*, 22 October 2001 373 (2001) (<http://www.parliament.the-stationery-office.co.uk/pa/cm200102/cmhansrd/cm011022/debindx/11022-x.htm>).
3. Det Norske Veritas *Assessment of Exposure to BSE Infectivity in the UK Sheep Flock* Report no. C782506 (The Meat and Livestock Commission, London, 1998).
4. Hadlow, W. J., Kennedy, R. C. & Race, R. E. Natural infection of Suffolk sheep with scrapie virus. *J. Infect. Dis.* **146**, 657–664 (1982).
5. van Keulen, L. J. M., Schreuder, B. E. C., Vromans, M. E. W., Langeveld, J. P. M. & Smits, M. A. Scrapie-associated prion protein in the gastrointestinal tract of sheep with natural scrapie. *J. Comp. Pathol.* **121**, 55–63 (1999).
6. van Keulen, L. J. M., Schreuder, B. E. C., Vromans, M. E. W., Langeveld, J. P. M. & Smits, M. A. Pathogenesis of natural scrapie in sheep. *Arch. Virol. Suppl.* **16**, 57–71 (2000).
7. Andreoletti, O. *et al.* Early accumulation of PrP^{Sc} in gut-associated lymphoid and nervous tissues of susceptible sheep from a Romanov flock with natural scrapie. *J. Gen. Virol.* **81**, 3115–3126 (2000).
8. Foster, J. D., Parnham, D. W., Hunter, N. & Bruce, M. Distribution of the prion protein in sheep terminally infected with BSE following experimental oral transmission. *J. Gen. Virol.* **82**, 2319–2326 (2001).
9. Jeffrey, M. *et al.* Oral inoculation of sheep with the agent of bovine spongiform encephalopathy (BSE). 1. Onset and distribution of disease-specific PrP accumulation in brain and viscera. *J. Comp. Pathol.* **124**, 280–289 (2001).
10. Fraser, H., Bruce, M. E., Chree, A., McConnell, I. & Wells, G. A. H. Transmission of bovine spongiform encephalopathy and scrapie to mice. *J. Gen. Virol.* **73**, 1891–1897 (1992).
11. Wells, G. A. H. *et al.* Infectivity in the ileum of cattle challenged orally with bovine spongiform encephalopathy. *Vet. Rec.* **135**, 40–41 (1994).
12. Hart, R. J., Church, P. N., Kempster, A. J. & Matthews, K. R. *Audit of Bovine and Ovine Slaughter and By-products Sector (Ruminant Products Audit)* (Leatherhead Food Research Association, London, 1997).
13. Hunter, N. PrP genetics in sheep and the implications for scrapie and BSE. *Trends Microbiol.* **5**, 331–334 (1997).
14. Hunter, N., Goldmann, W., Marshall, E. & O'Neill, G. Sheep and goats: natural and experimental TSEs and factors influencing incidence of disease. *Arch. Virol. Suppl.* **16**, 181–188 (2000).
15. Jeffrey, M. *et al.* Frequency and tissue distribution of infection-specific PrP in tissues of clinical scrapie and cull sheep obtained from scrapie affected farms in Shetland. *J. Comp. Pathol.* (submitted).
16. Baylis, M., Houston, E., Goldmann, W., Hunter, N. & McLean, A. R. The signature of scrapie: differences in the PrP genotype profile of scrapie-affected and scrapie-free UK sheep flocks. *Proc. R. Soc. Lond. B* **267**, 2029–2035 (2000).
17. Hunter, N. *et al.* Is scrapie solely a genetic disease? *Nature* **386**, 137 (1997).
18. Anderson, R. M. *et al.* Transmission dynamics and epidemiology of BSE in British cattle. *Nature* **382**, 779–788 (1996).
19. Ferguson, N. M., Donnelly, C. A., Woolhouse, M. E. J. & Anderson, R. M. The epidemiology of BSE in cattle herds in Great Britain II: Model construction and analysis of transmission dynamics. *Phil. Trans. R. Soc. Lond. B* **352**, 803–838 (1997).
20. Donnelly, C. A. & Ferguson, N. M. *Statistical Aspects of BSE and vCJD—Models for Epidemics* (Chapman & Hall and CRC, London, 2000).
21. Woolhouse, M. E. J. *et al.* Population dynamics of scrapie in a sheep flock. *Phil. Trans. R. Soc. Lond. B* **354**, 751–756 (1999).
22. Gravenor, M. B., Cox, D. R., Hoinville, L. J., Hoek, A. & McLean, A. R. The flock-to-flock force of infection for scrapie in Britain. *Proc. R. Soc. Lond. B* **268**, 587–592 (2001).
23. Hoinville, L. J. A review of the epidemiology of scrapie in sheep. *Rev. Sci. Techn. Office Int. Epizootics* **15**, 827–852 (1996).
24. Hoinville, L. J., Hoek, A., Gravenor, M. B. & McLean, A. R. Descriptive epidemiology of scrapie in Great Britain: results of a postal survey. *Vet. Rec.* **146**, 455–461 (2000).
25. Ghani, A. C., Ferguson, N. M., Donnelly, C. A. & Anderson, R. M. Predicted vCJD mortality in Great Britain. *Nature* **406**, 583–584 (2000).
26. d'Aignaux, J. N. H., Cousins, S. N. & Smith, P. G. Predictability of the UK variant Creutzfeldt–Jakob disease epidemic. *Science* **294**, 1729–1731; published online 25 October 2001 (10.1126/science.1064748).
27. Valleron, A.-J., Boelle, P.-Y., Will, R. & Ceberson, J.-Y. Estimation of epidemic size and incubation time based on age characteristics of vCJD in the United Kingdom. *Science* **294**, 1726–1728 (2001).
28. Hill, A. F. *et al.* Molecular screening of sheep for bovine spongiform encephalopathy. *Neurosci. Lett.* **255**, 159–162 (1998).
29. Maissen, M., Roedel, E., Glatzel, M., Goldmann, W. & Aguzzi, A. Plasminogen binds to disease-associated prion protein of multiple species. *Lancet* **357**, 2026–2028 (2001).
30. Hagenaars, T. J., Ferguson, N. M., Donnelly, C. A. & Anderson, R. M. Persistence patterns of scrapie in a sheep flock. *Epidemiol. Infect.* **127**, 157–167 (2001).

Supplementary Information accompanies the paper on *Nature's* website (<http://www.nature.com>).

Acknowledgements

This work was funded by the Food Standards Agency. N.M.F. and A.C.G. acknowledge funding from The Royal Society. C.A.D., T.J.H. and R.M.A. acknowledge funding from the Wellcome Trust. We are very grateful to M. Bruce, S. Bellworthy and M. Jeffrey for access to pre-publication data on infectivity. We also thank L. Hoinville, J. Wilesmith and R. Will for access to data. We thank N. Hunter, L. Green, J. Anderson, A. James, A. Bromley, H. Mason, P. Comer and members of the Spongiform Encephalopathy Advisory Committee for discussions.

Competing interests statement

The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to N.M.F. (e-mail: neil.ferguson@ic.ac.uk).

Reputation helps solve the 'tragedy of the commons'

Manfred Milinski, Dirk Semmann & Hans-Jürgen Krambeck

Department of Evolutionary Ecology, Max Planck Institute of Limnology, 24306 Plön, Postfach 165, Germany

The problem of sustaining a public resource that everybody is free to overuse—the 'tragedy of the commons'^{1–7}—emerges in many social dilemmas, such as our inability to sustain the global climate. Public goods experiments⁴, which are used to study this type of problem, usually confirm that the collective benefit will not be produced. Because individuals and countries often participate in several social games simultaneously, the interaction of these games may provide a sophisticated way by which to maintain the public resource. Indirect reciprocity⁸, 'give and you shall receive', is built on reputation and can sustain a high level of cooperation, as shown by game theorists^{9–11}. Here we show, through alternating rounds of public goods and indirect reciprocity games, that the need to maintain reputation for indirect reciprocity maintains contributions to the public good at an unexpectedly high level. But if rounds of indirect reciprocation are not expected, then contributions to the public good drop quickly to zero. Alternating the games leads to higher profits for all players. As reputation may be a currency that is valid in many social games, our approach could be used to test social dilemmas for their solubility.

Since Hardin¹ first described the 'tragedy of the commons', this type of social dilemma has been studied extensively by political and social scientists, economists and evolutionary theorists (see refs 2–7). Many of the experiments that have been carried out are a variant of the standard design⁴. In this model, four students seated at a table are each given an endowment of £5. They are then told that they can each choose to invest some or all of their £5 in a group project by putting, without discussion, an amount between £0 and £5 in an envelope. The experimenter will collect the 'contributions', total them up, double the amount, and then divide this money among the group.

The economic/game-theory prediction is that no one will ever contribute anything because each £1 contributed yields only £0.50 to its contributor, no matter what the others do. This is a public goods problem because the group would be best off (taking home £10 each) if all contributed £5. But individual self-interest is at odds with group interest. Usually people cooperate more than is predicted by standard economic theory⁴; however, observed cooperation is heterogeneous and declines over time (for example, see ref. 12). It has been shown that direct punishment of non-cooperators can cause a rise in the level of the average contribution to the public good^{13–15}, and cooperators are even prepared to pay a cost for punishing ('altruistic punishment')¹⁶.

We present an alternative way to maintain potentially a high level of contribution to the public good. It can be achieved through interaction with a second game that promises rewards for those with a good reputation in the public goods game. Theorists have shown that cooperation through indirect reciprocity can evolve^{9–11}. For indirect reciprocity, individuals who have helped others are given support, whereby the supporter builds up reputation^{8,17} or a positive image score^{9,10}. Experimental studies have confirmed that human subjects preferentially help others who have a positive image score^{18–20}. As players would risk their reputation if they would not cooperate in a public goods game that is alternated with the indirect reciprocity game, we predicted that alternating rounds of these two games would induce continuous cooperation in the public goods game, in contrast to a situation in which all public goods rounds were played first.

We tested these predictions with 114 first-year students who participated in 19 groups of 6 subjects each in a computerized experiment. The six subjects of each group could see a public screen on which instructions and the actual game was projected. They were told, first, that each person had a starting account of DM 20 (£10) and could gain or lose money dependent on his/her and the participants' decisions; second, that all decisions were anonymous and each player would be assigned a pseudonym (that is, a new identity) for the whole game; and last, that they would play in two different situations, an 'indirect reciprocity game' and a 'public goods game'.

Ten groups played one round of indirect reciprocity in which each subject was a potential donor once and a potential receiver once, and then one round of public goods. This alternating pattern was continued until round 16, thereafter four rounds of public goods were played. Every second group was told in round 17 that from then on only public goods rounds would follow until the end of the game. Nine other groups played eight rounds of public goods, followed by eight rounds of indirect reciprocity, followed by four rounds of public goods. Again, every second group was told in round 17 that from then on only public goods rounds would follow until the end of the game. In each round of an indirect reciprocity game, each potential receiver's history of giving both in the indirect reciprocity and the public goods game was displayed simultaneously for all players.

In groups that started with eight rounds of the public goods game initial cooperation declined as is usual in this type of game from round one to round eight (paired *t*-test between first and eighth round of public goods game, $n = 9$ groups, $t = 6.958$, $P < 0.0001$; Fig. 1). During the subsequent eight rounds of pairwise indirect reciprocity, cooperation was instantaneously re-established (comparison between eighth round of public goods game and first round of indirect reciprocity: paired *t*-test, $n = 9$ groups, $t = 2.9$, $P < 0.02$; to avoid pseudoreplication we use each group of six subjects as our statistical unit throughout this paper; all probabilities are two-tailed).

But in groups that started with one round of indirect reciprocity, followed by one round of public goods, and so on until round 16, the initial high cooperation level of the public goods game did not decline during the eight rounds of the public goods game (comparison between the first and the eighth round of public goods game; paired *t*-test, $n = 10$ groups, $t = 0.897$, $P = 0.40$), and was on average considerably higher than the cooperation level of the nine groups

that had started with eight rounds of public goods (unpaired *t*-test, d.f. = 17, $t = 4.83$, $P < 0.0002$; Fig. 1). When public goods and indirect reciprocity rounds were alternated, the public goods game elicited significantly more cooperation than the indirect reciprocity game (comparison between average cooperation of eight rounds public goods and eight rounds indirect reciprocity; paired *t*-test, $n = 10$ groups, $t = 3.99$, $P < 0.004$).

The high cooperation level in the public goods game was probably maintained in the following way. Players might have withheld help in the pairwise indirect reciprocity game from players who had refused to give in the preceding public goods round. The probability of receiving 'no' in the indirect reciprocity game was significantly higher for players that had refused to give in the preceding public goods round than for those who had given (Fig. 2a). Similarly, we found a positive correlation between the probability of receiving 'no' in the first round of the indirect reciprocity game and the rate of refused help during the block of eight rounds of the public goods game (mean Spearman's r per group = 0.49; s.e.m. = 0.15; Wilcoxon test against 0, $n = 9$ groups, $z = -2.1$, $P < 0.04$).

The hypothesis that interaction with the indirect reciprocity game keeps up cooperation in the public goods game is directly tested by the four rounds of public goods that groups in both treatments played in rounds 17–20. Every second group was told in round 17 that from then on only public goods rounds would follow until the end of the game. In these groups cooperation declined during the four public goods rounds, whereas cooperation was maintained when the risk of further rounds of indirect reciprocity was not excluded (Fig. 1; comparison of mean cooperation level during four rounds of public goods between five groups 'with' and five groups 'without announcement' after the alternating treatment, d.f. = 8,

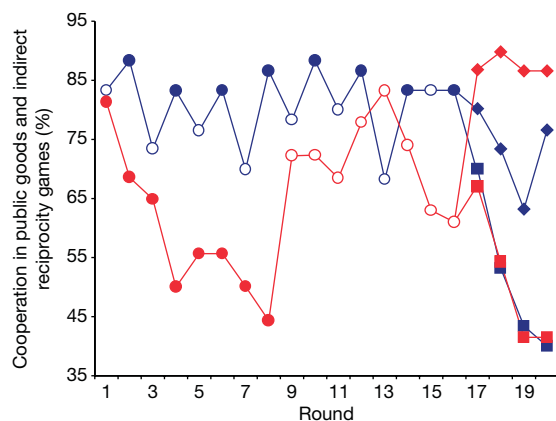


Figure 1 Percentage of cooperation ('yes') per group of six subjects in each round of the public goods game (filled symbols) and in each round of the indirect reciprocity game (open symbols). In one treatment, the groups alternated between rounds of indirect reciprocity and rounds of public goods until round 16 (blue); in the other treatment, groups started with eight consecutive rounds of the public goods game and continued with eight rounds of the indirect reciprocity game (red); in rounds 17–20, groups of both treatments played the public goods game, which was either announced, 'from now on only this type of game until the end' (squares), or not announced (diamonds).

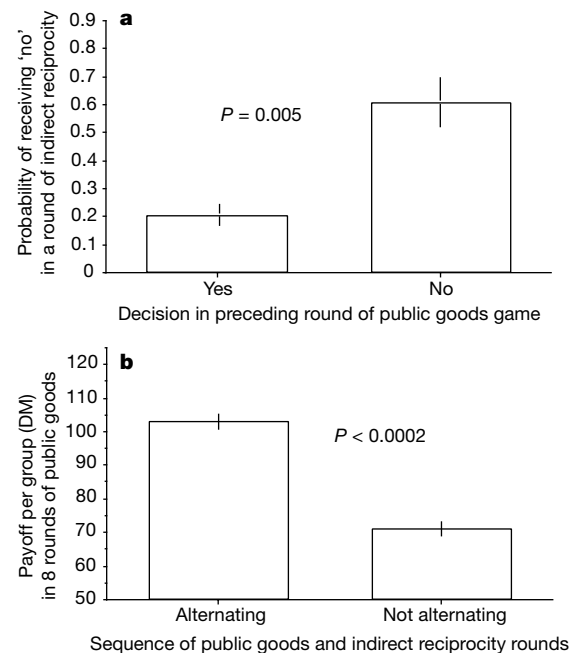


Figure 2 Consequences of cooperation in the public goods game. **a**, Probability of receiving 'no' in a round of the indirect reciprocity game depending on whether a subject had given either 'yes' or 'no' in the preceding round of the public goods game in the alternating treatment. The probability per group is shown (mean $\pm$ s.e.m.) for both situations; all individual situations were taken to generate one mean value of either type for each group; paired *t*-test with arcsine-transformed data, $n = 10$ groups, $t = 3.7$, $P = 0.005$. **b**, Payoff (DM) per group (mean $\pm$ s.e.m.) in the first 8 rounds of the public goods game of all groups that either alternated the indirect reciprocity and the public goods game during the first 16 rounds or started with 8 consecutive rounds of the public goods game and thereafter played 8 rounds of indirect reciprocity; unpaired *t*-test, d.f. = 17, $t = 4.83$, $P < 0.0002$.

unpaired t -test, $t = 4.456$, $P = 0.002$, and between five groups 'with' and four groups 'without announcement' after the block treatment, d.f. = 7, unpaired t -test, $t = 6.631$, $P = 0.0003$). Thus, the pending risk of further rounds of indirect reciprocity prevented cooperation in the public goods game from declining at least over four consecutive rounds.

Obviously, refusing to give in the public goods game reduced the reputation of a player to a similar extent as if this person had refused to give in the indirect reciprocity game: his potential donor in the next round of indirect reciprocity just followed the rules for indirect reciprocity and refused to give to someone with a low image score. This is different from punishing because it does not need any special punishing rule or motivation, and the potential donor actually saves money by refusing to give. A recent theoretical analysis²¹ suggests that reputation is essential for fostering social behaviour among selfish agents, which is confirmed experimentally here. The inclusion of reputation effects in the corresponding dynamical models leads to the evolution of economically productive behaviour, with agents contributing to the public good and either punishing those who do not or rewarding those who do²¹. Providing help in the indirect reciprocity game is a form of reward.

Cooperation in the public goods game paid off. Groups that alternated rounds of indirect reciprocity and public goods games, and thus were more cooperative in the public goods game, earned significantly more money during the first eight rounds of the public goods game than did groups that played the two games in blocks of eight rounds each (Fig. 2b). This shows that the 'tragedy of the commons' was no longer a tragedy; instead, the commons became productive and could be harvested. Two people usually interact in more than one situation, therefore their actions in one context may influence actions in another²². Many social dilemmas are a type of public goods game⁶, others have been identified as a type of indirect reciprocity game⁴. It therefore seems likely that the kind of interaction that we have staged experimentally occurs naturally in our society. There might be hidden social dilemmas that would show up only if the interaction with another game were removed. □

Methods

Indirect reciprocity game

Players were anonymous; each subject was assigned a pseudonym by the computer for the whole session of 20 rounds so that at any time, players could make their decisions contingent on the history of the game up to that time; each player knew his/her name but did not know who had been assigned the other names; the subjects were separated by opaque partitions and communicated their decisions with silent (piezo) switches; they knew that they would obtain their money after the game in a way that did not disclose their anonymity.

For the 'indirect reciprocity game'²⁰, each person was assigned repeatedly as either a potential donor or a potential receiver. For example, a potential donor, say 'Telesto', was asked on the public screen whether he would give to 'Galatea'. Telesto would lose DM 2.50 from his account and Galatea would gain DM 4 on her account if Telesto decided 'yes'. Telesto's decision (yes or no) was displayed for 2 s on the public screen. Everybody knew about the contributions of all players, for example, whether Galatea had given in previous rounds when he/she had been playing as the potential donor. The subjects also knew that there would be no direct reciprocity; if A was the potential donor of B, B would never be the potential donor of A. In each round of the indirect reciprocity game, each of the six players was once a potential donor and once a potential receiver.

Public goods game

For the 'public goods game'⁴, all six players were asked simultaneously whether they would contribute DM 2.50 to the public pool, the contents of which would then be doubled and redistributed evenly among all players irrespective of whether they had contributed. After all players had decided, each player's decision (yes or no), his/her contribution (that is, DM 2.50 or 0), and his/her gain (for example, DM 4.17 if all but one had contributed), was displayed below the pseudonyms for 20 s.

Received 8 November; accepted 4 December 2001.

- Hardin, G. The tragedy of the commons. *Science* **162**, 1243–1248 (1968).
- Dawes, R. Social dilemmas. *Annu. Rev. Psychol.* **31**, 169–193 (1980).
- Berkes, F., Feeny, D., McCay, B. J. & Acheson, J. M. The benefits of the commons. *Nature* **340**, 91–93 (1989).
- Ledyard, J. O. in *Handbook of Experimental Economics* (eds Kagel, J. H. & Roth, A. E.) 111–194 (Princeton Univ. Press, Princeton, 1995).
- Hardin, G. Extensions of "the tragedy of the commons". *Science* **280**, 682–683 (1998).

- Ostrom, E., Burger, J., Field, C. B., Norgaard, R. B. & Policansky, D. Sustainability—revisiting the commons: local lessons, global challenges. *Science* **284**, 278–282 (1999).
- Ostrom, E. *Governing the Commons* (Cambridge Univ. Press, Cambridge, 1999).
- Alexander, R. D. *The Biology of Moral Systems* (de Gruyter, New York, 1987).
- Nowak, M. A. & Sigmund, K. Evolution of indirect reciprocity by image scoring. *Nature* **393**, 573–577 (1998).
- Nowak, M. A. & Sigmund, K. The dynamics of indirect reciprocity. *J. Theor. Biol.* **194**, 561–574 (1998).
- Lotem, A., Fishman, M. A. & Stone, L. Evolution of cooperation between individuals. *Nature* **400**, 226–227 (1999).
- Fischbacher, U., Gächter, S. & Fehr, E. Are people conditionally cooperative? Evidence from a public goods experiment. *Econ. Lett.* **71**, 397–404 (2001).
- Boyd, R. & Richerson, P. J. Punishment allows the evolution of cooperation (or anything else) in sizable groups. *Ethol. Sociobiol.* **13**, 171–195 (1992).
- Fehr, E. & Gächter, S. Cooperation and punishment in public goods experiments. *Am. Econ. Rev.* **90**, 980–994 (2000).
- Gintis, H. *Game Theory Evolving* (Princeton Univ. Press, Princeton, 2000).
- Fehr, E. & Gächter, S. Altruistic punishment in humans. *Nature* **415**, 137–140 (2002).
- Zahavi, A. in *Cooperative Breeding in Birds: Long Term Studies in Behaviour and Ecology* (eds Stacey, P. B. & Koenig, W. D.) 105–130 (Cambridge Univ. Press, Cambridge, 1991).
- Wedekind, C. & Milinski, M. Cooperation through image scoring in humans. *Science* **288**, 850–852 (2000).
- Seinen, I. & Schram, A. Social status and group norms: indirect reciprocity in a helping experiment. Discussion Paper TI2001-003/1 (Tinbergen Institute, Amsterdam, 2001).
- Milinski, M., Semmann, D., Bakker, T. C. M. & Krambeck, H.-J. Cooperation through indirect reciprocity: image scoring or standing strategy? *Proc. R. Soc. Lond. B* **268**, 2495–2501 (2001).
- Sigmund, K., Hauert, C. & Nowak, M. A. Reward and punishment. *Proc. Natl Acad. Sci. USA* **98**, 10757–10762 (2001).
- Coleman, J. S. *Foundations of Social Theory* (Harvard Univ. Press, Cambridge, MA, 1990).

Acknowledgements

We thank students at Hamburg University for their participation; E. Heinz and W. Kiesewetter for support; E. Fehr and K. Sigmund for comments; and the Max Planck Institute of Meteorology at Hamburg for hospitality.

Competing interests statement

The authors declare that they have no competing financial interests.

Correspondence should be addressed to M.M.
(e-mail: milinski@mpil-ploen.mpg.de).

Species diversity enhances ecosystem functioning through interspecific facilitation

Bradley J. Cardinale, Margaret A. Palmer & Scott L. Collins

The University of Maryland, Department of Biology, College Park, Maryland 20742, USA

Facilitation between species is thought to be a key mechanism by which biodiversity affects the rates of resource use that govern the efficiency and productivity of ecosystems^{1–4}; however, there is no direct empirical evidence to support this hypothesis. Here we show that increasing the species diversity of a functional group of aquatic organisms induces facilitative interactions, leading to non-additive changes in resource consumption. We increased the richness and evenness of suspension-feeding caddisfly larvae (Insecta, Trichoptera) in stream mesocosms and found that the increased topographical complexity of the benthic habitat alters patterns of near-bed flow such that the feeding success of individuals is enhanced. Species diversity reduces 'current shading' (that is, the deceleration of flow from upstream to downstream neighbours), allowing diverse assemblages to capture a greater fraction of suspended resources than is caught by any species monoculture. The fundamental nature of this form of hydrodynamic facilitation suggests that it is broadly applicable to freshwater and marine habitats; in addition, it has several analogues in terrestrial ecosystems where fluxes of energy and matter can be

influenced by biophysical complexity^{3,5,6}. Thus, changes in species diversity may alter the probability of positive species interactions, resulting in disproportionately large changes in the functioning of ecosystems.

Rapid rates of species extinction and homogenization of the world's biota provide compelling reasons for determining how changes in biodiversity might affect the functioning of ecosystems^{7–9}. Current ecological theory predicts that species diversity can influence the consumption of resources that govern ecosystem processes through two types of effects: a 'complementarity effect', which occurs through either resource partitioning or facilitative interactions between species^{2,10,11}, and a 'selection effect', which occurs whenever species diversity is correlated with the chance of resource use being dominated by a single, productive taxon^{12,13}.

To date, research has focused primarily on how selection effects and complementarity through resource partitioning influence the functioning of terrestrial assemblages of plants⁸. But evidence showing that positive species interactions are pervasive in nature has led to speculation that interspecific facilitation may be a key mechanism by which biodiversity enhances the performance of ecosystems^{2,3,14}. Although research has shown that facilitative interactions between species of different functional groups can influence ecological processes^{11,15,16}, currently there is no mechanistic evidence that links species diversity in a functional group to positive species interactions that affect resource capture.

We studied a group of suspension feeders that are common in streams throughout the world (Trichoptera, Hydropsychidae).

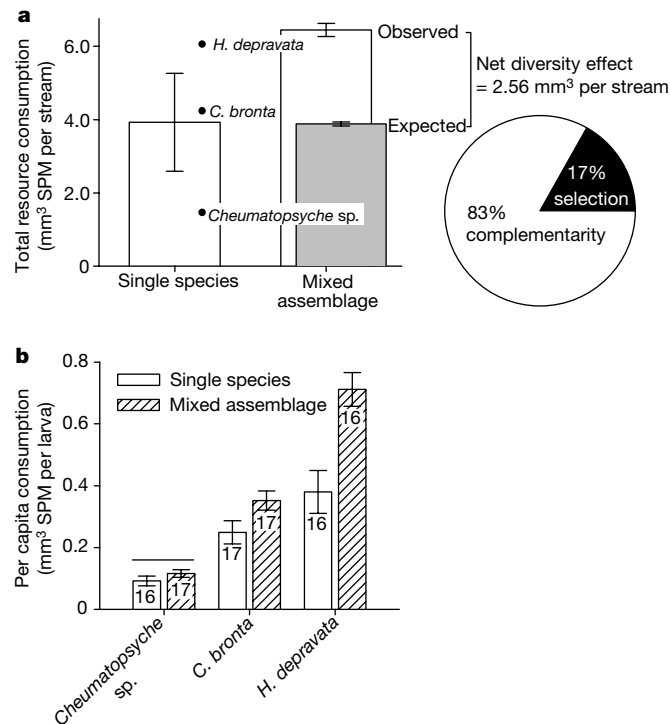


Figure 1 Effect of species diversity on resource consumption. **a**, Observed consumption of suspended particulate matter (SPM) in single-species and mixed-assemblage streams (open bars) with expected values (shaded bar) calculated from species performances in monoculture (data points). Bars indicate the mean $\pm$ s.e. ($n = 3$ streams). Inset shows that the net diversity effect (observed minus expected resource consumption) was caused primarily by complementarity. **b**, Enhanced per capita consumption of taxa in mixed assemblages owing to interspecific facilitation. Bars indicate the mean $\pm$ s.e.; those not connected are significantly different ($P < 0.05$). Numbers are the final abundance of species in monoculture (open bars) or the summed number of larvae in the three mixed-assemblage streams (hatched bars).

Larval hydropsychid caddisflies construct silk catchnets in the pore spaces of a streambed and passively filter suspended particulate matter (SPM) from the water. We focused on a system of three species that occur together throughout the eastern United States (*Hydropsyche depravata*, *Ceratopsyche bronta* and *Cheumatopsyche* sp.). This simple system allows the identification and generalization of mechanisms, but still represents local levels of diversity that are common for this and other aquatic functional groups¹⁷.

In aquatic habitats, biogenic structures are known to generate topographical 'roughness' (that is, spatial variation in surface features), which influences patterns of near-bed water flow and food delivery¹⁸. Because the physiognomy of feeding structures varies between species, we thought that increased diversity of hydropsychid caddisflies would increase bed roughness in a manner that alters near-bed flow and enhances the capture of organic resources. To test this hypothesis, we established two treatments of species diversity in stream mesocosms: 'single-species' streams (three streams, one per taxa) were colonized with 18 larvae of each caddisfly species; and 'mixed assemblages' (three streams) were established with 6 larvae of each taxa. This design resulted in a simultaneous increase in species richness and species evenness—the two aspects of species diversity. After allowing animals time to construct catchnets, we measured resource consumption by individual larvae and related this to the variation in near-bed velocity and bed roughness.

In streams with a mixed assemblage of species the total consumption of SPM was 66% greater, on average, than in single-species streams (Mann–Whitney $U = 9.00$, $P = 0.05$), and exceeded the total consumption in all species monocultures (Fig. 1a). The

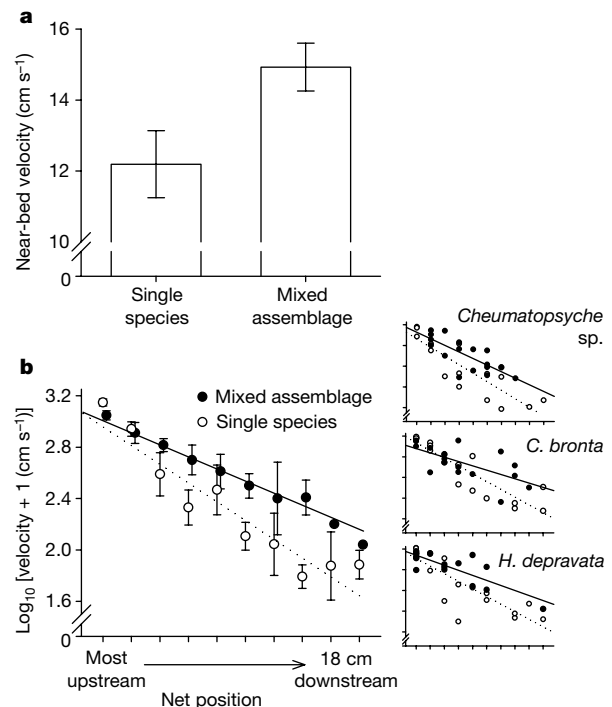


Figure 2 Effect of species diversity on flow. **a**, Near-bed current velocity measured at the entrance of caddisfly nets in single-species and mixed-assemblage streams (bars indicate the mean $\pm$ s.e. for 50 nets; means differ at $P = 0.02$). **b**, Upstream to downstream deceleration of flow into aggregations of suspension feeders. A significant interaction between treatment and net position (analysis of variance on log₁₀-transformed data, $F_{1,96} = 7.41$, $P = 0.01$) indicates that 'current shading' by upstream neighbours is more pronounced in single-species streams. Inset shows that there are consistent differences in current shading between diversity treatments for all taxa (axes are identical to those in main plot).

observed values of total SPM consumption in mixed assemblages was greater than that expected under the null hypothesis of no selection or complementarity effects² (paired $t = 11.41$, d.f. = 2, $P = 0.01$); thus, there was a significant positive net effect of species diversity on resource consumption (Fig. 1a). We partitioned the net diversity effect into two additive components due to selection or to complementarity². We found that only 0.43 mm³ of the increase in SPM consumption above expected values (17% of the net diversity effect) was caused by selection (species-specific) effects. The remaining 2.13 mm³ (83%) was the result of species complementarity, which can result from either resource partitioning or interspecific facilitation of resource capture^{2,4,14}.

In a previous study of this experimental system¹⁹, we showed that increasing the diversity of suspension feeders leads to an increase in the amount of SPM removed from stream water when the resource is not limiting, but diversity has no effect on SPM concentrations once the resource is depleted to limiting levels¹⁹. The latter result is contrary to what is expected from resource partitioning; that is, limiting resource concentrations should vary with species diversity when taxa differentially use a resource^{10,19}.

By contrast, the results of our current study are consistent with what is expected if suspension feeder diversity increases SPM consumption through interspecific facilitation of resource capture. We found that the feeding performance of individual larvae of each species was enhanced in the presence of other taxa. Per capita consumption of SPM by *H. depravata* and *C. bronta* increased by 87 and 41%, respectively, in mixed assemblages (*H. depravata*, $t = 3.76$, d.f. = 30, $P < 0.01$; *C. bronta*, $t = 2.13$, d.f. = 32, $P = 0.04$; Fig. 1b). Larval *Cheumatopsyche* also consumed 33% more SPM when inhabiting a mixed assemblage, but this increase was not significantly greater when compared with consumption by conspecifics in the single-species stream ($t = 1.20$, d.f. = 31, $P = 0.24$).

The hypothesis that increased resource consumption in mixed assemblages results from facilitative interactions is supported further by observed relationships between larval feeding performances and near-bed hydrodynamics. The delivery of food to suspension feeders is often proportional to flow velocity, particularly for passive filterers, such as hydropsychid caddisflies, that rely on water currents for the delivery of food particles²⁰. We found that per capita consumption of SPM was positively correlated with near-bed velocity (measured within millimetres in front of the entrance to catchnets) for both the single-species ($r = 0.64$, $P < 0.01$, $n = 49$ larvae) and mixed-assemblage streams ($r = 0.50$, $P < 0.01$, $n = 50$ larvae). There were positive correlations between velocity and per capita consumption for each individual species in monoculture ($P < 0.01$ for all), and for *H. depravata* ($P = 0.02$) and *C. bronta* ($P < 0.01$) in mixed assemblages. These results indicate that increased resource capture in mixed assemblages resulted from alterations in near-bed flow. Indeed, we found that velocity measured at the entrance of caddisfly nets averaged 22% faster in mixed assemblages (Mann–Whitney $U = 1592$, $P = 0.02$; Fig. 2a).

Differences in mean near-bed velocity between treatments resulted from differences in the magnitude of ‘current shading’ (that is, the blocking of current from upstream to downstream neighbours, Fig. 2b). Whereas flow tended to decelerate from upstream to downstream neighbours in both systems, near-bed current in front of catchnets diminished by an average -1.93 cm s^{-1} per downstream net site when species were alone, as compared with -1.47 cm s^{-1} in mixed assemblages. This trend was consistent for all three taxa (Fig. 2b), and as a result the average larva in mixed assemblages experienced a current that was 2.73 cm s^{-1} faster (22%) than when it was in monoculture.

Near-bed current velocity is known to be influenced by the density, height, spatial arrangement and topographical complexity of bed roughness elements^{18,21–25}. Density, height and spatial arrangement of caddisfly nets did not contribute to differences in

current shading between the treatments. We standardized larval densities at the beginning of the experiment and found no difference between treatments at the end of the study ($t < 0.01$, d.f. = 4, $P = 1.00$). Although there was interspecific variation in net heights (mean was 6.54, 6.36 and 4.2 mm for *H. depravata*, *C. bronta* and *Cheumatopsyche* sp., respectively), the mean height of nets for the three replicate single-species streams was equal to the mean height of nets in mixed assemblages ($t = 0.04$, d.f. = 4, $P = 0.97$). In addition, spatial aggregation of catchnets (the mean distance between neighbouring nets) was equivalent in the two treatments at the end of the experiment ($t = 1.44$, d.f. = 4, $P = 0.22$).

The only significant factor contributing to differences in bed roughness was topographical complexity (that is, the point-to-point variation in catchnet size measured as the standard deviation of net surface area). On average, topographical complexity was more than twofold greater in mixed assemblages than in single-species streams (s.d. = 18.61 versus 8.45 mm², respectively, $t = 6.41$, d.f. = 4, $P < 0.01$). Thus, placing taxa together led to streambeds characterized by non-uniformly sized elements of roughness. This is noteworthy because irregularly sized surface features can alter flow within an aggregation by inducing downstream eddies that increase diffusivity and turbulent mixing of water between the bed and overlying water column^{21,23–25}. Our data indicate that altered patterns of near-bed flow in mixed assemblages may result from greater topographical complexity caused by variation in the morphology of suspension-feeding structures.

Our study shows that increasing the species diversity of a group of aquatic arthropods leads to interspecific facilitation, and thus diverse assemblages can outperform species monocultures. The form of facilitation that we describe—where individuals influence the delivery of resources to neighbours through biophysical interactions—is likely to be widespread in aquatic and terrestrial systems. Current shading is common in freshwater and marine environments^{22,25–28}, and similar phenomena occur in terrestrial plant canopies and soil communities in which individuals affect vertical and horizontal fluxes of resources (such as gases, water and nutrients) to neighbours^{3,5,6}. The well known influence of structural complexity on fluid dynamics (both air and water)^{18,23,29,30} suggests, however, that biophysical effects of diversity should modify resource shading, and therefore the rates of resource capture. It has been proposed³, for example, that a positive relationship between bryophyte diversity and productivity results because a greater complexity of vertical structure helps to ‘trap’ water and facilitate plant survival during drought. Such differential impacts of species on the physical environment might be important mechanisms by which biodiversity generates positive interactions that enhance ecosystem functioning. □

Methods

Experimental design

The experiment was performed in recirculating stream mesocosms (0.1 m wide $\times$ 0.1 m deep $\times$ 1.0 m long) scaled to dimensions that maintain a dynamic similarity to conditions experienced by suspension feeders¹⁹. The working section of the streams (0.7 m downstream of the stream entrance) had 30 interstitial pores (3.6 mm deep $\times$ 15.0 mm long depressions such as found on rock surfaces) arranged in upstream to downstream rows. We released larvae uniformly over the surface of the working sections at the start of the experiment and allowed them to search freely for net sites. After 30 min of acclimation, the mid-channel flow was raised to $24 \pm 1 \text{ cm s}^{-1}$ (mean $\pm$ s.d.). Larvae that drifted were captured and replaced with new individuals so that 18 animals initially colonized all streams. Larvae were given 7 d to construct catchnets, during which we added 46 mg of 63–149- μm suspended particulate matter per day to streams as food, set lighting to a 16:8 h day:night cycle, and maintained a constant water temperature (18–19 °C).

Near-bed flow

After 7 d, a Sontek 10-MHz acoustic doppler velocimeter (ADV) was used to record a 2-min time series (at 10 Hz, 1,200 readings) of three-dimensional velocity 5–6 mm above the streambed and 5–7 mm upstream of the entrance of every caddisfly net. To avoid acoustic interference, the sampling cell of the ADV was compressed to a nominal 4 mm in

height using software available from Sontek. From each time series we calculated mean near-bed velocity independent of flow direction. Mean near-bed velocity was compared between treatments using a non-parametric Mann–Whitney *U*-test because variances could not be transformed to satisfy parametric assumptions.

Resource consumption

After measuring near-bed flow, 278 mg of SPM stained with Rose Bengal dye was released as a single pulse into each stream¹⁹. Larvae were allowed to feed for 15 min (a duration less than gut passage times) before they were removed from their nets and frozen. We dissected larval guts later and measured the diameter and band length of stained SPM in foreguts using a dissecting microscope and ocular micrometer. Because foreguts are essentially cylindrical, the consumption of SPM by each larva was calculated as $\text{mm}^3 \text{ SPM by } \pi \times \text{band length in foregut} \times (1/2 \text{ foregut diameter})^2$. Per capita consumption was compared between treatments using *t*-tests. Total resource consumption (the summed consumption of SPM by all larvae inhabiting a stream) was compared between treatments using a non-parametric Mann–Whitney *U*-test because variances could not be transformed to satisfy parametric assumptions. We used a paired *t*-test to compare observed resource consumption in mixed assemblages with the total expected SPM consumption².

Bed roughness

At the end of the experiment we recorded the downstream location of every caddisfly net and measured their maximum heights and widths. We calculated the average maximum height and density of the roughness elements, as well as their aggregation and topographical complexity. Aggregation measures the spacing between roughness elements as the mean euclidian distance between neighbouring nets. Topographical complexity measures the spatial uniformity or non-uniformity of roughness elements as the standard deviation of the parabolic area (in mm^2) of catchnets. An s.d. of 0 indicates a uniform streambed (no topographical complexity), whereas a higher s.d. indicates greater streambed complexity. We compared all four aspects of bed roughness between treatments using *t*-tests.

Received 8 October; accepted 26 November 2001.

- Hector, A. *et al.* Plant diversity and productivity experiments in European grasslands. *Science* **286**, 1123–1127 (1999).
- Loreau, M. & Hector, A. Partitioning selection and complementarity in biodiversity experiments. *Nature* **412**, 72–76 (2001).
- Mulder, C., Uliassi, D. & Doak, D. Physical stress and diversity–productivity relationships: The role of positive interactions. *Proc. Natl Acad. Sci. USA* **98**, 6704–6708 (2001).
- Tilman, D. *et al.* Diversity and productivity in a long-term grassland experiment. *Science* **294**, 843–845 (2001).
- Jones, C. G., Lawton, J. H. & Shachack, M. Positive and negative effects of organisms as physical ecosystem engineers. *Ecology* **78**, 1946–1957 (1997).
- Lutge, U. *Physiological Ecology of Tropical Plants* (Springer, Berlin, 1997).
- Naeem, S., Thompson, L. J., Lawler, S. P., Lawton, J. H. & Woodfin, R. M. Declining biodiversity can alter the performance of ecosystems. *Nature* **368**, 734–737 (1994).
- Cardinale, B. J., Nelson, K. & Palmer, M. A. Linking species diversity to the functioning of ecosystems: on the importance of environmental context. *Oikos* **91**, 175–183 (2000).
- Chapin, F. S. *et al.* Consequences of changing biodiversity. *Nature* **405**, 234–242 (2000).
- Tilman, D., Lehman, D. & Thompson, K. Plant diversity and ecosystem productivity: theoretical considerations. *Proc. Natl Acad. Sci. USA* **94**, 1857–1861 (1997).
- Hooper, D. & Vitousek, P. The effects of plant composition and diversity on ecosystem processes. *Science* **277**, 1302–1305 (1997).
- Huston, M. Hidden treatments in ecological experiments: re-evaluating the ecosystem function of biodiversity. *Oecologia* **110**, 449–460 (1997).
- Wardle, D. A. Is “sampling effect” a problem for experiments investigating biodiversity–ecosystem function relationships? *Oikos* **87**, 403–410 (1999).
- Fridley, J. D. The influence of species diversity on ecosystem productivity: how, where, and why? *Oikos* **93**, 514–526 (2001).
- Tilman, D. *et al.* The influence of functional diversity and composition on ecosystem processes. *Science* **277**, 1300–1302 (1997).
- Chapin, S. *et al.* Ecosystem consequences of changing biodiversity. *BioScience* **48**, 45–52 (1998).
- Palmer, M. A. *et al.* Biodiversity and ecosystem function in freshwater sediments. *Ambio* **26**, 571–577 (1997).
- Nowell, A. R. M. & Jumars, P. Flow environments of aquatic benthos. *Annu. Rev. Ecol. Syst.* **15**, 303–328 (1984).
- Cardinale, B. J. & Palmer, M. A. Disturbance moderates biodiversity–ecosystem function relationships: evidence from caddisfly assemblages in stream mesocosms. *Ecology* (in the press).
- Loudon, C. & Alstad, D. N. Theoretical mechanisms of particle capture: Predictions for hydropsychid distributional ecology. *Am. Nat.* **135**, 360–381 (1990).
- Eckman, J. E., Nowell, A. R. M. & Jumars, P. J. Sediment destabilization by animal tubes. *J. Mar. Res.* **39**, 361–374 (1981).
- Johnson, A. Flow around phoronids: Consequences of a neighbor to suspension feeders. *Limnol. Oceanogr.* **35**, 1395–1401 (1990).
- Huetzel, M. & Gust, G. Impact of bioirrigation on interfacial solute exchange in permeable sediments. *Mar. Ecol. Prog. Ser.* **89**, 253–267 (1992).
- Butman, C. A., Frechette, M., Geyer, W. R. & Starczak, V. R. Flume experiments on food supply to the blue mussel *Mytilus edulis* L. as a function of boundary-layer flow. *Limnol. Oceanogr.* **39**, 1755–1768 (1994).
- Sebens, K. P., Witting, J. & Helmuth, B. Effects of water flow and branch spacing on particle capture by the reef coral *Madracis mirabilis* (Duchassaing and Michelotti). *J. Exp. Mar. Biol. Ecol.* **211**, 1–28 (1997).

- Hart, D. D. The adaptive significance of territoriality in filter-feeding larval blackflies (Diptera: Simuliidae). *Oikos* **46**, 88–92 (1986).
- Englund, G. Asymmetric resource competition in a filter-feeding stream insect (*Hydropsyche siltalai*: Trichoptera). *Freshwat. Biol.* **26**, 425–432 (1991).
- Okamura, B. Microhabitat variation and patterns of colony growth and feeding in a marine bryozoan. *Ecology* **73**, 1502–1513 (1992).
- Vogel, S. *Life in Moving Fluids* (Princeton Univ. Press, Princeton, 1994).
- Finnigan, J. Turbulence in plant canopies. *Annu. Rev. Fluid Mech.* **32**, 519–571 (2000).

Acknowledgements

We thank D. Doak, D. Hart, M. Loreau, P. Morin, S. Naeem, K. Sebens, D. Tilman, J. Thomson and T. Welnitz for comments; and S. Brooks for advice on hydrodynamic measurements. This work was supported by grants from the National Science Foundation to M.A.P. and to B.J.C.

Competing interests statement

The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to B.J.C. (e-mail: bc84@umail.umd.edu).

Humans integrate visual and haptic information in a statistically optimal fashion

Marc O. Ernst* & Martin S. Banks

Vision Science Program/School of Optometry, University of California, Berkeley 94720-2020, USA

When a person looks at an object while exploring it with their hand, vision and touch both provide information for estimating the properties of the object. Vision frequently dominates the integrated visual–haptic percept, for example when judging size, shape or position^{1–3}, but in some circumstances the percept is clearly affected by haptics^{4–7}. Here we propose that a general principle, which minimizes variance in the final estimate, determines the degree to which vision or haptics dominates. This principle is realized by using maximum-likelihood estimation^{8–15} to combine the inputs. To investigate cue combination quantitatively, we first measured the variances associated with visual and haptic estimation of height. We then used these measurements to construct a maximum-likelihood integrator. This model behaved very similarly to humans in a visual–haptic task. Thus, the nervous system seems to combine visual and haptic information in a fashion that is similar to a maximum-likelihood integrator. Visual dominance occurs when the variance associated with visual estimation is lower than that associated with haptic estimation.

The estimate of an environmental property by a sensory system can be represented by

$$\hat{S}_i = f_i(S) \quad (1)$$

where S is the physical property being estimated and f is the operation by which the nervous system does the estimation. The subscripts refer to the modality (i could also refer to different cues within a modality). Each estimate, $\hat{S}_i$, is corrupted by noise. If the noises are independent and gaussian with variance σ_i^2 , and the bayesian prior is uniform, then the maximum-likelihood estimate

* Present address: Max Planck Institute for Biological Cybernetics, Tübingen 72076, Germany.

(MLE) of the environmental property is given by

$$\hat{S} = \sum_i w_i \hat{S}_i \quad \text{with} \quad w_i = \frac{1/\sigma_i^2}{\sum_j 1/\sigma_j^2} \quad (2)$$

Thus, the MLE rule states that the optimal means of estimation (in the sense of producing the lowest-variance estimate) is to add the sensor estimates weighted by their normalized reciprocal variances^{8–15}. If the MLE rule is used to combine visual and haptic estimates, $\hat{S}_V$ and $\hat{S}_H$, the variance of the final (visual–haptic) estimate, $\hat{S}$, is

$$\sigma_{VH}^2 = \frac{\sigma_V^2 \sigma_H^2}{\sigma_V^2 + \sigma_H^2} \quad (3)$$

Thus, the final estimate has lower variance than either the visual or the haptic estimator. Implementation of MLE integration is shown for two hypothetical cases in Fig. 1.

We examined visual–haptic integration quantitatively to determine whether human performance is optimal. Observers looked at and/or felt a raised ridge (Fig. 2) and judged its height (vertical extent). To work out the predictions of the MLE rule, we first determined the variances of the visual and haptic height estimates (within-modality) by conducting discrimination experiments. In the haptic-alone experiment, observers indicated which of two sequentially presented ridges was taller from haptic information alone; in the visual-alone experiment, they did the same from visual information alone. There were four conditions in the visual experiment that differed in the amount of noise in the stimulus (see Methods). By adding noise we made the visually specified height less reliable.

Visual-alone and haptic-alone discrimination data are shown in Fig. 3a. The proportion of trials in which the observer indicated that the comparison stimulus (variable height) appeared taller than the standard stimulus (fixed height of 55 mm) is plotted as a function of

the height of the comparison stimulus. The dashed red line and symbols represent the haptic discrimination data, and the solid blue curves with open symbols represent the visual data for the four levels of noise. These psychometric functions were well fit by cumulative gaussian functions.

The discrimination threshold is defined as the difference between the point of subjective equality (PSE) and the height of the comparison stimulus when it is judged taller than the standard stimulus 84% of the time. The 84% point corresponds to $\sqrt{2}$ times the standard deviation of the underlying estimator. The haptic discrimination threshold was roughly 0.085 times the average ridge height (which was 55 mm). As the noise increased from 0 to 200%, the visual discrimination thresholds increased from 0.04 to 0.2 times the average height. Thus, when the visual noise was 0%, the visual discrimination threshold was roughly half the haptic threshold; when the visual noise was 200%, the visual threshold was more than double the haptic threshold.

In the visual–haptic experiment, observers simultaneously looked at and felt two raised ridges that were presented sequentially. In one presentation the visually and haptically specified heights were equal (comparison stimulus); in the other presentation they differed (standard stimulus). The difference (Δ) in the specified heights was ± 6 , ± 3 or 0 mm (the average of S_H and S_V was 55 mm). For each Δ in the standard stimulus (randomly presented), the height of the comparison stimulus was varied randomly from trial to trial (47–63 mm). On each trial, the observer indicated which stimulus seemed taller. Figure 3b shows the proportion of trials in which the comparison stimulus was chosen as taller as a function of the height of the comparison stimulus. From these psychometric functions, we estimated the PSE—the comparison height appearing equal to the standard height—and the just-discriminable change in height (threshold).

Using the within-modality data, we can predict what an observer using MLE will do when presented visual and haptic information

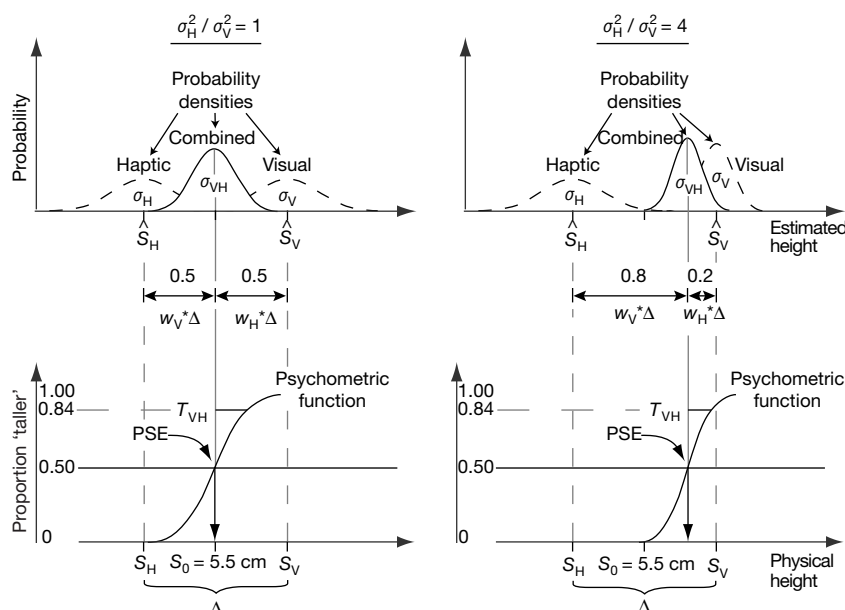


Figure 1 Maximum-likelihood estimation integration: two hypothetical situations. Visually and haptically specified heights differ by Δ . Dashed gaussians in the top panels represent probability densities of the (unbiased) estimated height from visual and haptic assessment, and solid gaussians represent probability densities for the combined estimate. On the left, the visual and haptic variances are equal ($\sigma_V^2/\sigma_H^2 = 1$) and both their weights are 0.5 (equation (2)). The mean of the combined probability density is therefore equal to the mean of the visual and haptic densities and the variance is reduced by half (equation (3)). If the observer bases judgements of relative height on the combined

probability density, the psychometric function would be a cumulative gaussian (bottom left) with a point of subjective equality (PSE) equal to the average of the visual and haptic heights of the standard stimulus. On the right, the haptic variance is four times the visual variance: $\sigma_H^2/\sigma_V^2 = 4$. By equation (2), the visual weight (w_V) is 0.8 and the haptic weight (w_H) is 0.2. Thus, the combined probability density is shifted towards the visual estimate and its variance is 0.8 times the visual variance (equation (3)). Accordingly, the psychometric function should be shifted so that the PSE is closer to the visual height of the standard stimulus.

simultaneously, and we compare these predictions to the performance in the visual–haptic experiment. First, we describe the analysis of the PSE data and predictions for the weights. From equation (2) and the relationship between threshold and estimator variance:

$$\frac{w_V}{w_H} = \frac{\sigma_H^2}{\sigma_V^2} = \frac{T_H^2}{T_V^2} \quad (4)$$

where T_H and T_V are the haptic and visual thresholds (84% points in Fig. 3a). Incorporating the normalization assumption ($w_V + w_H = 1$), the predicted weights for optimal integration are

$$w_V = \frac{T_H^2}{T_V^2 + T_H^2} \quad \text{and} \quad w_H = \frac{T_V^2}{T_V^2 + T_H^2} \quad (5)$$

The predicted visual weights are represented by the curve and shaded surround in Fig. 3c. The predicted weights vary significantly with the amount of visual noise in the stimulus: the visual weights are higher when the noise level is low, and lower when the noise level is high. Assuming that the visual and haptic estimators are on average unbiased ($\hat{S}_V = S_V$ and $\hat{S}_H = S_H$), the weights can be derived experimentally:

$$w_V = (\text{PSE} - S_H)/(S_V - S_H) \quad (6)$$

where PSE is the height of the comparison stimulus that matched the apparent height of the standard stimulus. The visually and haptically specified heights in the standard stimulus, S_V and S_H , are indicated on the right ordinate. Figure 3c shows that as the noise level was increased the visual weight decreased, and the PSE shifted from S_V towards S_H . Because the noise level varied randomly from trial to trial, the weights must have been set within the 1-s stimulus presentation. Below, we suggest a mechanism for such dynamic weight adjustment. In summary, the predicted and observed PSEs are similar, suggesting that humans do combine visual and haptic information in a fashion similar to MLE integration.

According to the MLE rule, the combined estimates should have lower variance, and therefore lower discrimination thresholds, than either the visual or haptic estimate alone (equation (3)). To derive predictions for the visual–haptic discrimination thresholds, we rewrite equation (3):

$$T_{VH}^2 = \frac{T_V^2 T_H^2}{T_V^2 + T_H^2} \Leftrightarrow \frac{1}{T_{VH}^2} = \frac{1}{T_V^2} + \frac{1}{T_H^2} \quad (7)$$

The predicted and observed thresholds are shown in Fig. 3d. The open symbols represent the visual-alone thresholds and the dashed line represents the haptic-alone threshold. The shaded area represents the predicted visual–haptic thresholds; they are always lower than the visual-alone and haptic-alone thresholds at the corresponding noise level. The filled purple symbols represent the observed visual–haptic discrimination thresholds; as noise level increases, the just-noticeable difference in height becomes greater. Most notably, the predicted and observed visual–haptic discrimination thresholds are similar. As with the PSE data, this indicates that human observers may combine visual and haptic information in a manner similar to MLE integration.

In summary, we found that height judgements were remarkably similar to those predicted by the MLE integrator. Thus, the nervous system seems to combine visual and haptic information in a fashion similar to the MLE rule: visual and haptic estimates are weighted according to their reciprocal variances (equation (2)). Naturally, it is important to determine whether this rule characterizes the estimation of other stimulus properties such as shape, depth, localization, roughness or compliance.

The relative contributions of vision and haptics to perceiving such object properties have been studied previously. For example, subjects have grasped a square while looking at it through a distorting lens that made it appear rectangular¹. The shape of the unified percept was determined almost completely by vision, so the phenomenon was called ‘visual capture’. Numerous studies have

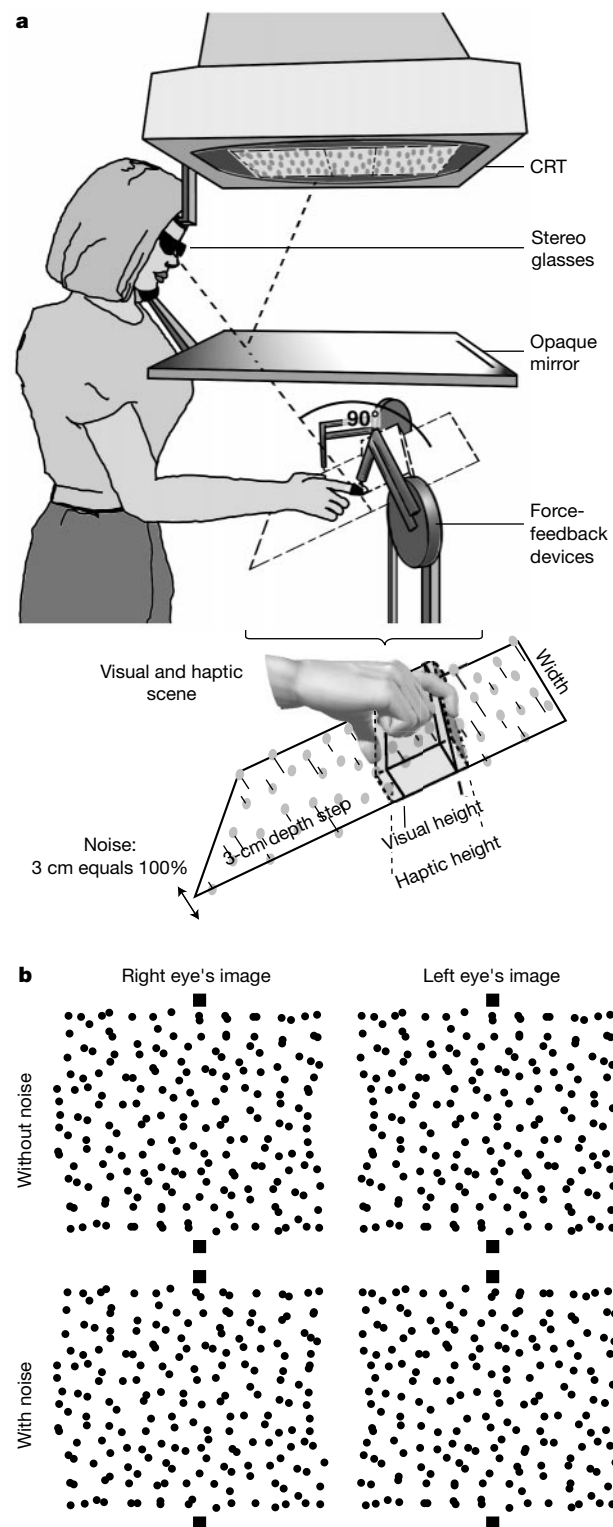


Figure 2 Apparatus and stimuli. **a**, Observers viewed the reflection of the visual stimulus, presented on a cathode ray tube (CRT) binocularly in a mirror. CrystalEyes (StereoGraphics) liquid-crystal shutter glasses were used to present binocular disparity. The surfaces of the stimuli were perpendicular to the line of sight. A head and chin rest limited head movements. The right hand was beneath the mirror and could not be seen. The haptic stimulus was presented with two PHANTOM force-feedback devices, one each for the index finger and thumb. **b**, Stereograms representative of visual stimuli, which should be viewed by cross-fusing. Top row, stereogram has no noise and the horizontal bar is raised above the background. Bottom row, stereogram of bar and background contains noise, with random displacements of dots parallel to the line of sight.

replicated visual capture in shape and size perception^{3,16,17}, depth perception¹⁸ and localization^{21,19–21}. But visual capture does not occur in the perception of surface roughness^{6,22}; instead, perceived roughness is affected nearly equally by haptics and vision. Does a dynamic cue-combination rule, such as the one described here, determine the degree to which vision or haptics dominates? The statistically optimal means of combining visual and haptic information—the MLE rule—predicts that visual capture should occur whenever the visual estimate of a property has much less variance than that of the haptic estimate. Haptic capture should be observed when the reverse occurs. We observed behaviour like visual capture when the visual stimulus was noise-free, and behaviour similar to haptic capture when the visual stimulus was quite noisy (Fig. 3c).

In visual–haptic tasks, the MLE integrator described here always uses information from both sensory systems, so the combined percept will always reflect both sources of information. Different behaviour may be observed when the discrepancy between two information sources is large. With large discrepancies between information sources, the nervous system may exhibit robust behaviour¹⁰ in which a discrepant source is discounted^{9,23}. In robust estimation, the weights associated with different information sources are determined by more than the variances of the sensory estimators; they are also determined by the discrepancy between their estimates. In our experiments, the differences between the visually and haptically specified stimuli were never greater than 11%, and the visual–haptic data always exhibited an influence of both sensory systems. If the differences had been larger, we might

have observed discounting of one sense. It would be interesting to know whether such vetoing occurs when observers become aware of the conflict between visual and haptic inputs.

If the nervous system implements MLE integration, the weights must be proportional to the reciprocal variances of the probability densities associated with the visual and haptic estimates of the environmental property in question (equation (2) and Fig. 1). Of course, the variances change from one object property to the next (such as size, shape or roughness) and from one situation to another (for example, visual variance increases as the lighting is degraded). Does the nervous system need to calculate or learn the variances associated with the visual and haptic estimators for each property and situation to implement MLE integration? Although explicit calculation or learning may occur²⁴, there are plausible schemes in which explicit calculation of variances or weights is unnecessary.

Consider, for example, a population of visual and haptic neurons, each sensitive to a range of heights. Each neuron has a preferred height but also responds to other heights (that is, each neuron has a tuning function). If the visual input specifies the height clearly (that is, high contrast, noise-free, and so on), then the visual neurons preferring that height respond vigorously and those preferring other heights respond less, and the distribution of response across the population of visual neurons has a well defined peak. Assume that the distribution across the population of haptic neurons has a less well defined peak. Multiplication of these two distributions (point-by-point multiplication where the two populations are in registration according to the stimulus property being estimated)²⁵ yields a

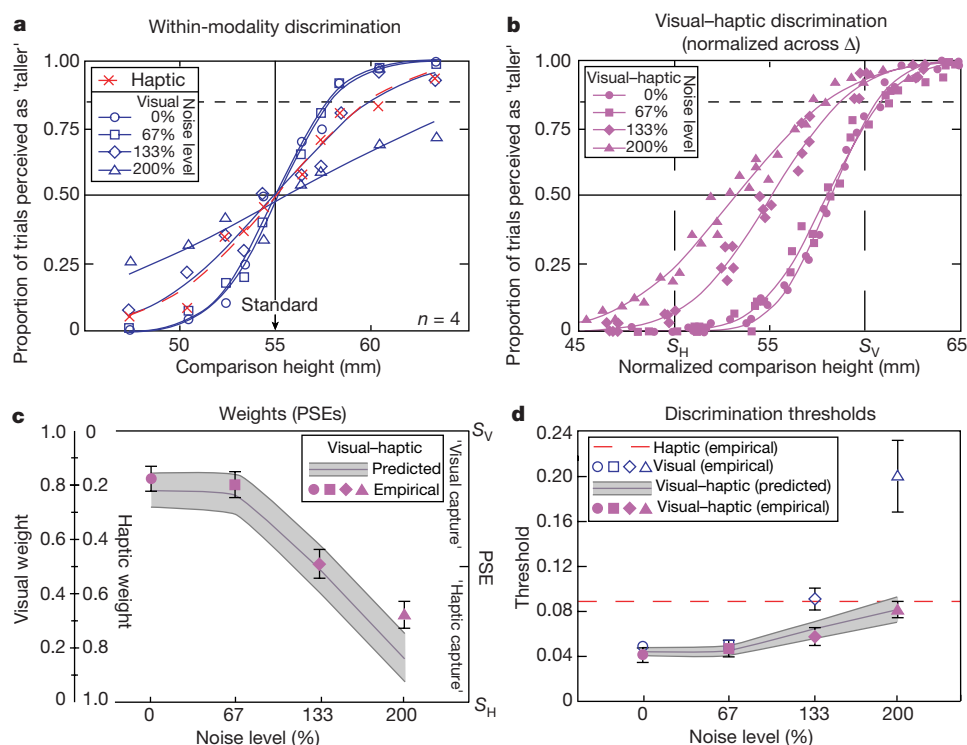


Figure 3 Predictions and experimental data. **a**, Within-modality discrimination. Proportion of trials in which a comparison was perceived as taller than the standard stimulus is plotted against the height of the comparison stimulus. Data were averaged for observers. Height of the standard stimulus was 55 mm. Haptic discrimination data are represented by red crosses and the dashed curve (best-fitting cumulative gaussian); visual discrimination data are represented by blue curves, which correspond to four noise levels. **b**, Visual–haptic discrimination. The average height of visual–haptic standard stimulus was 55 mm; the height difference, Δ , varied from -6 to $+6$ mm. To plot the data on the same coordinates, the psychometric functions for each Δ were shifted laterally by $w_V\Delta/2$ (w_V is obtained from the regression of PSE against Δ). Purple curves represent the different visual noise levels. **c**, Weights and PSEs. Abscissa represents the noise level, left

ordinate represents visual weight (w_V ; haptic weight is $1 - w_V$) and right ordinate represents the PSEs relative to S_V and S_H . Purple symbols represent observed visual weights obtained from regression analysis of PSEs (equation 6) across Δ . Shaded area represents predicted weights expected from within-modality discrimination (**a**; equation (5)); its height represents predicted errors given the standard errors of the within-modality discrimination. **d**, Combined and within-modality discrimination thresholds. Just-noticeable differences in height are plotted against noise level. Thresholds are taken from psychometric functions in **a** and **b**. Dashed red line represents haptic-alone threshold; open blue symbols represent visual-alone thresholds; filled purple symbols represent combined visual–haptic thresholds. Shaded area represents predicted visual–haptic thresholds (equation (7)).

peak response closer to the visual than the haptic peak, as with MLE integration. If degrading the visual input causes the response distribution of the visual neurons to spread, then multiplication of the visual and haptic distributions yields a peak closer to the haptic peak, again as in MLE integration. Thus, the estimator variances (and therefore the weights) do not have to be calculated explicitly: the behaviour of an MLE integrator might be achieved through interactions among populations of visual and haptic neurons. □

Methods

Stimuli

The stimulus was a horizontal bar raised 30 mm above a plane; the bar and plane were perpendicular to the line of sight (Fig. 2a). The width of the bar was 150 mm; its height varied but the average was $S_0 = 55$ mm. Observers viewed the bar binocularly and/or grasped it with the index finger and thumb in order to estimate its height (Fig. 2a). Its vertical position was varied randomly from trial to trial.

The haptic stimulus was generated using two PHANTOM (SensAble Technologies) force-feedback devices (Fig. 2a), one each for the index finger and thumb. Finger and thumb movements had all six degrees of freedom in the 20-cm³ workspace. The three-dimensional positions of the tips of the finger and thumb were monitored, and appropriate forces were applied when the tips reached the positions of the simulated haptic objects. PHANTOMS compellingly simulate haptic properties such as the size, shape and stiffness of the bar. The apparatus was calibrated to align the visual and haptic stimuli spatially.

The visual stimulus was a random-dot stereogram simulating the background plane and bar (Fig. 2). The dots subtended 8 arcmin at the 50-cm viewing distance. Dot density was roughly 9 dots per degree². New dots were displayed with each presentation. The positions of the finger and thumb (tracked by the PHANTOMS) were indicated by small three-dimensional markers; the markers were visible until the bar was touched. Noise was added to the visual display to vary its reliability. The noise was a random displacement of the dot depths in the stereogram (direction parallel to line of sight). The displacements were drawn from a random uniform distribution whose range was 0, 67, 133 or 200% of the 3-cm depth step between the bar and plane. The displacement of the dots is shown in Fig. 2. No noise was added to the haptic display.

We wanted the presentation times for the visual and haptic stimuli to be identical. In the vision-alone discrimination experiment, the standard and comparison stimuli were displayed for 1 s each. In the haptic-alone discrimination experiment, the haptic stimulus began when the thumb and finger both contacted the bar and ended after 1 s. In the visual-haptic trials, the visually specified bar did not appear until the bar was touched by both fingers simultaneously and the visual and haptic stimuli were extinguished after 1 s.

Procedure

Within-modality discrimination was measured in a two-interval, forced-choice scheme. Each trial consisted of the sequential (visual or haptic) presentation of two bars. In the standard interval, the bar was always 55 mm tall and in the comparison interval, it was shorter or taller than 55 mm. The standard and comparison stimuli were randomly assigned to the first or second interval. The observer indicated the interval containing the apparently taller stimulus. The comparison height was varied according to the method of constant stimuli. We plotted psychometric functions, that is the proportion of trials in which the comparison was perceived as taller than the standard against the comparison height. From these functions, we could determine PSEs and discrimination thresholds. Half of the vision-alone and haptic-alone discrimination experiments were conducted (in random order) before the visual-haptic experiment and the other half were conducted after.

In the visual-haptic experiment, observers were again presented two stimuli sequentially: the standard stimulus (visually and haptically specified heights differed by Δ) and the comparison stimulus (visually and haptically specified heights were the same). The average height of the standard stimulus was always 55 mm and the visual-haptic difference (Δ) ranged from -6 to +6 mm. The height of the comparison varied. The observer indicated the interval containing the apparently taller stimulus. Trials with different conflicts were presented in random order to prevent visuomotor adaptation.

No feedback was given in any of the experiments. Four observers (aged 22 to 33), naive to the experimental purpose, participated. They were right handed and had normal or corrected-to-normal vision. The observers were questioned at the end: none of them had noticed the conflicts between the visually and haptically specified heights.

Received 25 July; accepted 15 November 2001.

1. Rock, I. & Victor, J. Vision and touch: An experimentally created conflict between the two senses. *Science* **143**, 594–596 (1964).
2. Hay, J. C., Pick, H. L. Jr & Ikeda, K. Visual capture produced by prism spectacles. *Psychonomic Sci.* **2**, 215–216 (1965).
3. Warren, D. H. & Rossano, M. J. in *The Psychology of Touch* (eds Heller, M. A. & Schiff, W.) 119–137 (Erlbaum, Hillsdale, New Jersey, 1991).
4. Power, R. P. The dominance of touch by vision: Sometimes incomplete. *Perception* **9**, 457–466 (1980).
5. Welch, R. B. & Warren, D. H. Immediate perceptual response to intersensory discrepancy. *Psychol. Bull.* **88**, 638–667 (1980).

6. Lederman, S. J. & Abbott, S. G. Texture perception: Studies of intersensory organization using a discrepancy paradigm, and visual versus tactual psychophysics. *J. Exp. Psychol. Hum. Percept. Perform.* **7**, 902–915 (1981).
7. Heller, M. A. Haptic dominance in form perception with blurred vision. *Perception* **12**, 607–613 (1983).
8. Clark, J. J. & Yuille, A. L. *Data Fusion for Sensory Information Processing Systems* (Kluwer Academic, Boston, 1990).
9. Blake, A., Bülthoff, H. H. & Sheinberg, D. Shape from texture: Ideal observer and human psychophysics. *Vision Res.* **33**, 1723–1737 (1993).
10. Landy, M. S., Maloney, L. T., Johnston, E. B. & Young, M. Measurement and modeling of depth cue combination: In defense of weak fusion. *Vision Res.* **35**, 389–412 (1995).
11. Gharamani, Z., Wolpert, D. M. & Jordan, M. I. in *Self-organization, Computational Maps, and Motor Control* (eds Morasso, P. G. & Sanguineti, V.) 117–147 (Elsevier, Amsterdam, 1997).
12. Knill, D. C. Discrimination of planar surface slant from texture: Human and ideal observers compared. *Vision Res.* **38**, 1683–1697 (1998).
13. Backus, B. T. & Banks, M. S. Estimator reliability and distance scaling in stereoscopic slant perception. *Perception* **28**, 417–442 (1999).
14. van Beers, R. J., Sittig, A. C. & Denier van der Gon, J. J. Integration of proprioceptive and visual position information: An experimentally supported model. *J. Neurophysiol.* **81**, 1355–1364 (1999).
15. Schrater, P. R. & Kersten, D. How optimal depth cue integration depends on the task. *Int. J. Comp. Vis.* **40**, 71–89 (2000).
16. Gibson, J. J. Adaptation, after-effect, and contrast in the perception of curved lines. *J. Exp. Psychol.* **16**, 1–31 (1933).
17. Festinger, L., Burnham, C. A., Ono, H. & Bamber, D. Efference and the conscious experience of perception. *J. Exp. Psychol.* **74** (4), 1–36 (1967).
18. Singer, G. & Day, R. H. Visual capture of haptically judged depth. *Percept. Psychophys.* **5**, 315–316 (1969).
19. Tastevin, J. En partant de l'expérience d'Aristote. *L'Encephale* **1**, 57–84 (1937).
20. Mon-Williams, M., Wann, J. P., Jenkinson, M. & Rushton, K. Synaesthesia in the normal limb. *Proc. R. Soc. Lond. B* **264**, 1007–1010 (1997).
21. Pavani, F., Spence, C. & Driver, J. Visual capture of touch: out-of-the-body experiences with rubber gloves. *Psycholog. Sci.* **11**, 353–359 (2000).
22. Heller, M. A. Visual and tactual texture perception: Intersensory cooperation. *Percept. Psychophys.* **31**, 339–344 (1982).
23. Banks, M. S. & Backus, B. T. Extra-retinal and perspective cues cause the small range of the induced effect. *Vision Res.* **38**, 187–194 (1998).
24. Ernst, M. O., Banks, M. S. & Bülthoff, H. H. Touch can change visual slant perception. *Nature Neurosci.* **3**, 69–73 (2000).
25. Peña, J. L. & Konishi, M. Auditory spatial receptive fields created by multiplication. *Science* **292**, 249–252 (2001).

Acknowledgements

We thank M. Landy for comments on the manuscript; and H. Ernst, X. Moncada, C. Alderson and S. Kashiwada for participating as observers. This research was supported by research grants from Air Force Office of Scientific Research and the National Institutes of Health, and by an equipment grant from Silicon Graphics.

Competing interests statement

The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to M.O.E. (e-mail: marc.ernst@tuebingen.mpg.de).

Effects of grouping in contextual modulation

Michael H. Herzog* & Manfred Fahle*†

* Human Neurobiology, University of Bremen, Argonnenstrasse 3, 28211 Bremen, Germany

† City University, Department of Optometry and Visual Science, Tait Building, Northampton Square, London EC1V 0HB, UK

Perception of a visual target and the responses of cortical neurons can be strongly influenced by a context surrounding the target^{1–27}. This observation relates to the fundamental issue of how cortical neurons code objects of the external world. In high-contrast regimes, embedding a target in an iso-oriented context reduces neural responses and deteriorates performance in psychophysical experiments. Performance from orthogonal surrounds is better than that from iso-oriented ones^{1–17}. This contextual interference is often postulated to be caused by long- or short-range interac-

tions between neurons tuned to orientation. Here we show, using a new illusion called 'shine-through' as a sensitive psychophysical probe, that the orientation difference between target and context does not determine performance. Instead, contextual modulation depends on the overall spatial structure of the context. We propose that contextual suppression vanishes if the contextual elements are grouped to an independent and coherent object.

In the shine-through illusion^{4,5}, a briefly displayed vernier precedes a grating referred to as the 'standard grating', which comprises more than seven aligned elements (see Fig. 1). Subjectively, the vernier shines through the grating and appears brighter, wider, longer and superimposed on the grating. Vernier discrimination thresholds can be as small as 20 arcsec, whereas those for an unmasked vernier can be around 8 arcsec, indicating a masking effect of the grating. Homogeneity of the standard grating is a prerequisite for the effect to occur; even very subtle spatial deviations from this homogeneity markedly diminish the shine-through effect^{4,5}.

Four iso-oriented contextual elements markedly degrade performance (that is, the thresholds of offset discrimination increase; Fig. 2a, g). This finding agrees with most studies using high-contrast target regimes^{1–18}. Performance improves, however, if more iso-oriented contextual elements are displayed. For 25 contextual lines, performance is about the same as that for the standard grating, and significantly better than that for a display with 4 contextual elements (Fig. 2a, b). Horizontal contextual lines yield analogous results: performance improves if small horizontal elements become part of a long horizontal line (Fig. 2c, d). One line on both sides of the standard grating (Fig. 2e) yields significantly lower performance than two lines placed on only one side (Fig. 2f). Thus, the overall spatial structure rather than the orientation difference between target and context determines contextual suppression.

Both a decrease (Fig. 2f) and an increase (Fig. 2b) of the number of contextual lines improve performance compared with four contextual lines; therefore, the number of lines is neither necessary nor sufficient to predict performance. A better explanation is that contextual suppression vanishes if the elements of the context are grouped to a coherent object.

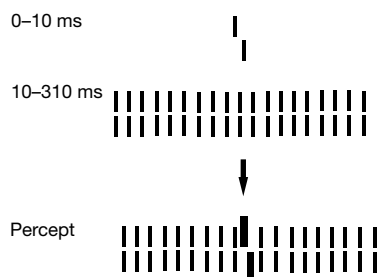


Figure 1 The shine-through effect. A vernier, consisting of two abutting lines, presented for a short time (10–30 ms) shines through a following grating if the grating comprises more than seven elements. The shine-through element appears to be wider, brighter and even longer than the preceding vernier really is. This grating is called the 'standard grating'. With less than seven grating elements the vernier remains invisible to the observer^{4,5}. Except for offset, all spatial parameters of grating elements were the same as those of the preceding vernier. The horizontal spacing between grating elements was 200 arcsec, and verniers were 1,260 arcsec long presented for the shortest duration at which shine-through just occurred. Stimuli were displayed on an analog monitor (HP 1334A) controlled by a Power Macintosh computer through fast 16-bit digital-to-analog converters (pixel rate 1 MHz). Luminance was about 80 cd m⁻². Subjects were asked to discriminate the offset direction (left versus right) of the shine-through element by pressing one of two push buttons. Incorrect responses were followed by an error signal produced by the computer. Thresholds were determined with an adaptive staircase method (the offset size of the vernier yielding 75% correct responses). Statistical analysis was performed on logarithmic data. See ref. 5.

The contextual gratings shown in Fig. 3a, which do not diminish shine through, strongly degrade performance if attached to the standard grating (Fig. 3b). Attaching these gratings creates an inhomogeneous central grating; therefore, vernier discrimination decreases (see ref. 5). Only a minor spatial change has occurred because the vertical gap between the contextual and the standard grating is only 200 arcsec wide; in other words, it is much smaller than the receptive fields of most cortical neurons. Moreover, the overall intensity of the whole context is identical. Thus, it seems to matter strongly whether the context is an independent part of the visual display (ref. 6, but see refs 15, 17).

Contextual interference emerges in a fast dynamic process. We presented single contextual verniers preceding two contextual gratings and a vernier preceding the central standard grating (Fig. 4b). The contextual verniers appear as distinct shine-through elements, whereas the visibility of the central target vernier is strongly diminished; therefore, vernier discrimination deteriorates markedly (Fig. 4e) compared to the condition of Fig. 4a. Arrays of more than one contextual vernier do not, in general, shine through a following

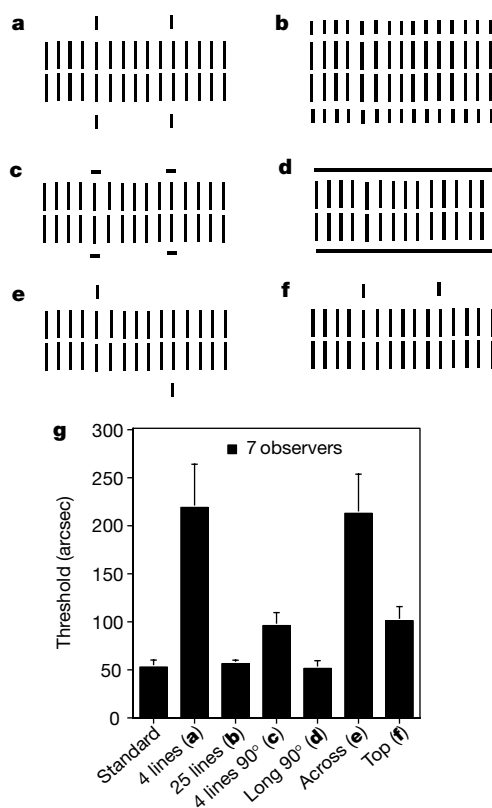


Figure 2 Contextual interference cannot be predicted by contextual orientation. A vernier (not shown) preceded a grating with 25 elements (15 are shown) flanked by contextual lines presented simultaneously with the grating. The vertical gap between contextual lines and grating was 200 arcsec. **a**, Four vertical lines of 400 arcsec length were displayed above and below the third grating element on both sides. **b**, Twenty-five contextual elements appeared above and below the standard grating. **c**, Four lines were oriented orthogonal to the grating (that is, horizontal). **d**, Two long horizontal lines flanked the whole grating. **e**, Two lines were displayed, one above left and the other below right of middle. **f**, Only two lines above the grating were displayed. **g**, Performance for 4 contextual lines is lower, hence thresholds are higher, than that for 25 elements, even though these 4 lines are part of the 25 contextual lines (paired *t*-test, $P = 0.0024$). For some observers, the shine-through element is completely invisible for four contextual lines and vernier discrimination thresholds cannot be determined (a threshold of 350 arcsec is recorded in these cases). Four small horizontal lines interfere more strongly than two long horizontal lines (paired *t*-test, $P = 0.0105$). If two lines are arranged on one side of the grating, performance is better than if elements are in corresponding positions on different sides of the grating (paired *t*-test, $P = 0.005$).

grating. Thus, three aligned contextual verniers, each preceding contextual gratings, remain invisible (Fig. 4c). Even so, the arrays of contextual verniers inhibit the single central vernier, which therefore remains largely invisible. Offset discrimination of this vernier deteriorates strongly—comparable to the condition with single contextual verniers. Thus, suppression by preceding contextual verniers is not caused by the subjective visibility of the preceding contextual elements themselves.

A vernier and a grating three times longer than standard (Fig. 4d), mimicking an attachment of contextual verniers and gratings to the central standard vernier and grating, improve performance significantly compared with the separated gratings and verniers shown in Fig. 4b. Attachment here leads to temporal and spatial homogeneity in contrast to the conditions in Fig. 3, where spatial inhomogeneities of the standard grating degrade performance. Again, contextual intensity is identical for stimuli in Fig. 4b and d, whereas thresholds are not.

Contextual suppression diminishes if the suppressive contextual elements of Fig. 2a are presented without a following grating (data not shown). Therefore, it is the backward masking nature of the shine-through effect that reveals contextual modulation during the first milliseconds of cortical information processing^{10,15,28}.

In most psychophysical and physiological studies using high-contrast targets, contextual suppression is strongest for iso-oriented contexts and decreases for orthogonal ones^{1–17}. For example, responses of a neuron to a line presented in its receptive field strongly decrease when several iso-oriented contextual lines appear in the surrounding non-classical receptive field^{1–3}. Neural responses increase if the number of these contextual lines decreases (ref. 1, but see ref. 2). The underlying neural mechanisms are thought to rely on orientation-specific horizontal connections or on high-level back projections. For low-contrast targets a more complex pattern occurs, revealing facilitating effects in addition^{8,16,19–23}.

Contextual surrounds are homogeneous in most studies. We compared various contextual layouts and thus could evaluate context–context modulation in addition to context–target modulation. Contrary to the above results, we found that increasing the number of iso-oriented contextual elements improves performance (see Fig. 2a, b). An extended iso-oriented, homogeneous surround and a long, orthogonal (that is, horizontal) contextual line yield performance comparable to that of the standard condition (that is, the standard grating without any surround). Clearly, commonly

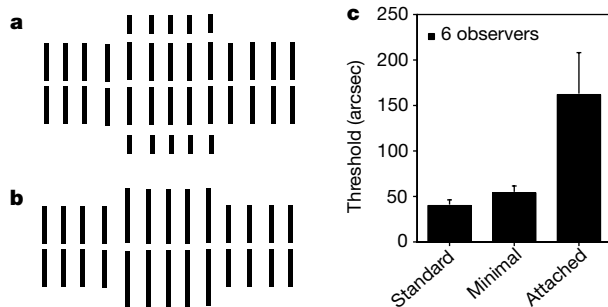


Figure 3 Independence of the context matters. **a**, We determined the minimal number of iso-oriented contextual elements for which interference is absent so that fewer elements would yield a deterioration of performance. For four observers this number was five lines, whereas for two subjects it was seven (only 13 elements of the standard grating are shown). The vertical gap between contextual lines and standard grating was 200 arcsec. **b**, Contextual gratings were directly attached to the standard grating; that is, the gap between contextual and standard gratings was removed. **c**, Performance strongly deteriorates if gratings that cause almost no contextual interference ('minimal') are attached to the central grating ('attached'; paired *t*-test, $P = 0.0085$). Note that the minimal gratings contain the four contextual elements from the condition in Fig. 2a. Vertical bars symbolize standard errors of the means of all six paid student observers.

used low-level descriptions such as orientation, number, and overall intensity of contextual elements are not sufficient to explain the results.

We propose that the overall spatial structure of the context, rather than its individual low-level features, determines modulation. A context does not interfere if it can be grouped to a coherent and independent gestalt—in other words, if the context is an independent entity and all its elements are bound together, or if unbound parts are remote from the target.

Although the influence of contextual elements seems to be complex if based on a low-level stimulus description, the neural mechanism underlying contextual grouping may be simple^{8–10,29}. For example, to explain the shine-through effect, a model has been proposed on the basis of a single neural layer with lateral interactions whose dynamics are described by a partial differential equation (C. W. Eurich, U. Ernst and M.H.H., unpublished data; see also ref. 29). According to this equation, neural activity representing the inner elements of a grating decreases compared with that at the edges. For small gratings, this enhanced activity corresponding to the edges suppresses the neural activity corresponding to the

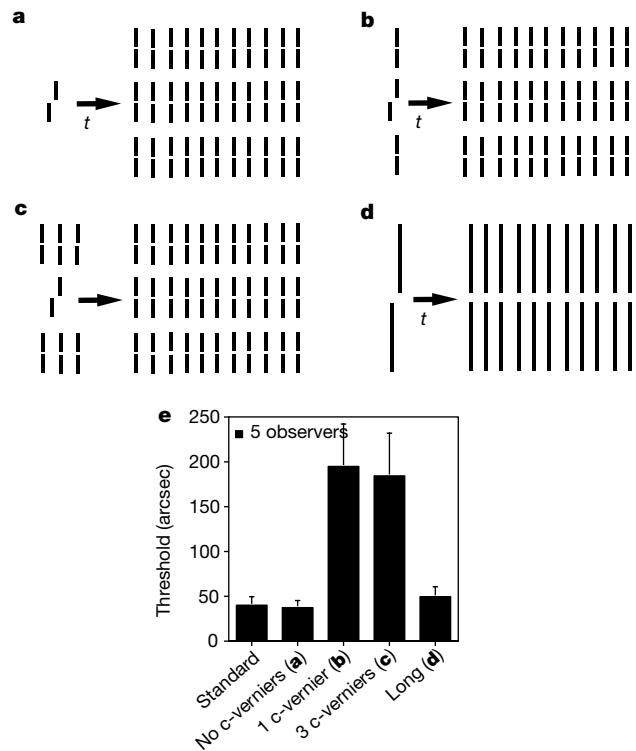


Figure 4 Shine through interferes with shine through. Here, time (*t*) proceeds from left to right rather than from top to bottom. Preceding verniers always appeared at the central position of the corresponding gratings. The standard condition was applied first (see Fig. 1). **a**, As in the condition with 25 contextual elements (Fig. 2b), a vernier preceded the standard grating that was flanked above and below by two identical copies of itself. **b**, As in **a**, except that the contextual gratings were each preceded by a straight vernier, and an offset vernier preceded the standard grating. **c**, Three aligned verniers preceded each of the contextual gratings, an offset one preceded the standard grating. **d**, The grating and vernier were three times longer than in the standard condition. This stimulus had the same energy as in condition **b** and almost the same size. **e**, The standard and the condition in **a** yield comparable performance levels, reaffirming that contextual energy *per se* is not important. Performance in **b** is greatly impaired compared with that in **a** (paired *t*-test, $P = 0.0053$); the contextual verniers (c-verniers), although visible themselves, seriously impair the visibility and therefore offset discrimination for the central vernier. The condition in **c** shows the same impairment as in **b**. Here, the triple verniers, although not visible, impair the discriminability of the central vernier. Contextual interference vanishes when the separated contextual gratings are replaced by a lengthening of the standard grating and vernier (paired *t*-test, $P = 0.0081$ for **c** versus **d**).

vernier. For the same reason, activity corresponding to four isolated (that is, unbound) contextual elements (Fig. 2a) might not be decreased by neighbouring contextual elements and, thus, might produce high neural activity inhibiting the vernier target. For the extended standard grating, shine through is possible because little neural activity occurs in the vernier's close neighbourhood. Analogously, extended contextual gratings might not interfere with the vernier because activity corresponding to their inner elements is weak. Therefore, context–context inhibition prevents context–target suppression of the vernier target.

Contextual elements might directly influence the firing pattern of cortical neurons that code the vernier. For example, the activity of a neuron that is responding vigorously to a vernier target should diminish when isolated contextual elements are presented together with the standard grating. By contrast, increasing the number of contextual elements should yield a rebound of activity. Because vernier display times are short, we expect strongest modulation effects in the transient response of the neurons—that is, in the activity immediately after the presentation of the target. □

Received 17 July; accepted 20 November 2001.

1. Knerim, J. J. & van Essen, D. C. Neuronal responses to static texture patterns in area V1 of the alert macaque monkey. *J. Neurophys.* **67**, 961–980 (1992).
2. Li, W., Thier, P. & Wehrhahn, C. Contextual influence on orientation discrimination of humans and responses of neurons in V1 of alert monkeys. *J. Neurophys.* **83**, 941–954 (2000).
3. Nothdurft, H.-C., Gallant, J. L. & van Essen, D. C. Response modulation by texture surround in primate area V1: Correlates of pop-out under anesthesia. *Vis. Neurosci.* **16**, 15–34 (1999).
4. Herzog, M. H. & Koch, C. Seeing properties of an invisible element: feature inheritance and shine-through. *Proc. Natl Acad. Sci. USA* **98**, 4271–4275 (2001).
5. Herzog, M. H., Fable, M. & Koch, C. Spatial aspects of object formation revealed by a new illusion, shine-through. *Vis. Res.* **41**, 2325–2335 (2001).
6. Mareschal, I., Sceniak, M. P. & Shapley, M. Contextual influences on orientation discrimination: binding local and global cues. *Vis. Res.* **41**, 1915–1930 (2001).
7. Wehrhahn, C., Li, W. & Westheimer, G. Patterns that impair discrimination of orientation in human vision. *Perception* **25**, 1053–1064 (1996).
8. Stemmler, M., Usher, M. & Niebur, E. Lateral interactions in primary visual cortex: a model bridging physiology and psychophysics. *Science* **269**, 1877–1880 (1995).
9. Li, Z. Contextual influences in V1 as a basis for pop out and asymmetry in visual search. *Proc. Natl Acad. Sci. USA* **96**, 10530–10535 (1999).
10. Wolfson, S. S. & Landy, M. S. Long range interactions between oriented texture elements. *Vis. Res.* **39**, 933–945 (1999).
11. Nothdurft, H. C. Texture segmentation and pop-out from orientation contrast. *Vis. Res.* **31**, 1073–1078 (1991).
12. Cannon, M. W. & Fullenkamp, S. C. Spatial interactions in apparent contrast: inhibitory effects among grating patterns of different spatial frequencies, spatial positions and orientations. *Vis. Res.* **31**, 1985–1998 (1991).
13. Lamme, V. A. The neurophysiology of figure-ground segregation in primary visual cortex. *J. Neurosci.* **15**, 1605–1615 (1995).
14. Zipser, K., Lamme, V. A. F. & Schiller, P. H. Contextual modulation in primary visual cortex. *J. Neurosci.* **16**, 7376–7389 (1996).
15. Caputo, G. The role of the background: Texture segregation and figure-ground segmentation. *Vis. Res.* **36**, 2815–2826 (1996).
16. Sengpiel, F., Sen, A. & Blakemore, C. Characteristics of surround inhibition in cat area 17. *Exp. Brain Res.* **116**, 216–228 (1997).
17. Sillito, A. M., Grieve, K. L., Jones, H. E., Cudeiro, J. & Davis, J. Visual cortical mechanisms detecting focal orientation discontinuities. *Nature* **378**, 492–496 (1995).
18. Macknik, S. L. & Livingstone, M. S. Neuronal correlates of visibility and invisibility in the primate visual system. *Nature Neurosci.* **1**, 144–149 (1998).
19. Polat, U. & Sagi, D. Lateral interactions between spatial channels: suppression and facilitation revealed by lateral masking experiments. *Vis. Res.* **33**, 993–999 (1993).
20. Kapadia, M. K., Ito, M., Gilbert, C. D. & Westheimer, G. Improvement in visual sensitivity by changes in local context: parallel studies in human observers and in V1 of alert monkeys. *Neuron* **15**, 843–856 (1995).
21. Levitt, J. B. & Lund, J. S. Contrast dependence of contextual effects in primate visual cortex. *Nature* **387**, 73–76 (1997).
22. Bonneh, Y. & Sagi, D. Effects of spatial configuration on contrast detection. *Vis. Res.* **38**, 3541–3543 (1998).
23. Yu, C. & Levi, D. Surround modulation in human vision unmasked by masking experiments. *Nature Neurosci.* **3**, 724–728 (2000).
24. Das, A. & Gilbert, C. D. Topography of contextual modulations mediated by short-range interactions in primary visual cortex. *Nature* **399**, 655–661 (1999).
25. Chubb, C., Sperling, G. & Solomon, J. A. Texture interactions determine perceived contrast. *Proc. Natl Acad. Sci. USA* **86**, 9631–9635 (1989).
26. Weissstein, N. & Harris, C. S. Visual detection of lines segments: an object-superiority effect. *Science* **186**, 752–755 (1974).
27. Banks, W. P. & White, H. Lateral interference and perceptual grouping in visual detection. *Percept. Psychophys.* **36**, 285–295 (1984).
28. Kurylo, D. D. Time course of perceptual grouping. *Percept. Psychophys.* **59**, 142–147 (1997).
29. Ernst, U. A., Pawelzik, K. R., Sahar-Pikielny, C. & Tsodyks, M. V. Intracortical origin of visual maps. *Nature Neurosci.* **4**, 431–436 (2001).

Acknowledgements

This work was supported by the Center of Excellence (SFB) 'Neurocognition' of the German Research Council (Deutsche Forschungsgemeinschaft). We thank M. Repnow for technical support; and B. Gillam, C. Eurich, U. Ernst, A. Etzold, B. Zenger, K. Neumann, F. Kandil, J. Solomon, M. Morgan, J. Zanker and U. Schmonsees for discussions.

Competing interests statement

The authors declare that they have no competing financial interests.

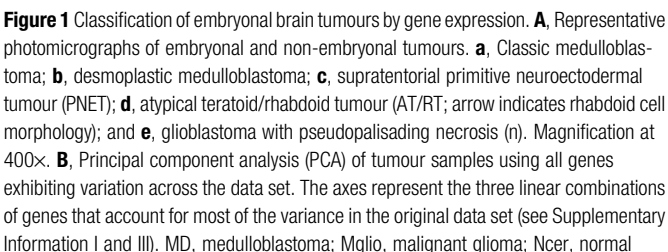
Correspondence and requests for materials should be addressed to M.H.H. (e-mail: mherzog@uni-bremen.de).

Prediction of central nervous system embryonal tumour outcome based on gene expression

Scott L. Pomeroy*, **Pablo Tamayo†**, **Michelle Gaassenbeek†**, **Lisa M. Sturla***, **Michael Angelo†**, **Margaret E. McLaughlin‡**, **John Y. H. Kim***, **Liliana C. Goumnerova||**, **Peter M. Black||**, **Ching Lau¶**, **Jeffrey C. Allen#**, **David Zagzag***, **James M. Olson***, **Tom Curran††**, **Cynthia Wetmore††**, **Jaclyn A. Biegel‡‡**, **Tomaso Poggio§§**, **Shayan Mukherjee§§**, **Ryan Rifkin§§**, **Andrea Califano|||**, **Gustavo Stolovitzky|||**, **David N. Louis¶¶**, **Jill P. Mesirov†**, **Eric S. Lander†††** & **Todd R. Golub†††**

* Division of Neuroscience, Department of Neurology; † Department of Pathology; ‡ Department of Pathology; § Department of Neurosurgery; ¶ Department of Medicine, Children's Hospital; § Department of Pediatric Oncology, Dana-Farber Cancer Institute; and ¶ Department of Pathology and Neurosurgical Service, Massachusetts General Hospital, Harvard Medical School, Boston, Massachusetts 02115, USA
† Whitehead Institute/MIT Center for Genome Research, and §§ McGovern Institute, Center for Biological and Computational Learning, AI Lab, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA
¶ Division of Pediatric Oncology, Baylor College of Medicine, Houston, Texas 77030, USA
Beth Israel Medical Center, New York 10128, USA
* Department of Pathology, New York University School of Medicine, New York 10016, USA
** Clinical Research Division, Fred Hutchinson Cancer Research Center, Seattle, Washington 98109, USA
†† Department of Developmental Neurobiology, St Jude Children's Research Hospital, Memphis, Tennessee 38105, USA
‡‡ Division of Human Genetics, The Children's Hospital of Philadelphia, Department of Pediatrics, University of Pennsylvania School of Medicine, Philadelphia, Pennsylvania 19104, USA
§§ Department of Biology, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA
||| IBM Watson Research Center, Yorktown Heights, New York 10598, USA

Embryonal tumours of the central nervous system (CNS) represent a heterogeneous group of tumours about which little is known biologically, and whose diagnosis, on the basis of morphologic appearance alone, is controversial. Medulloblastomas, for example, are the most common malignant brain tumour of childhood, but their pathogenesis is unknown, their relationship to other embryonal CNS tumours is debated^{1,2}, and patients' response to therapy is difficult to predict³. We approached these problems by developing a classification system based on DNA microarray gene expression data derived from 99 patient samples. Here we demonstrate that medulloblastomas are molecularly distinct from other brain tumours including primitive neuroectodermal tumours (PNETs), atypical teratoid/rhabdoid tumours



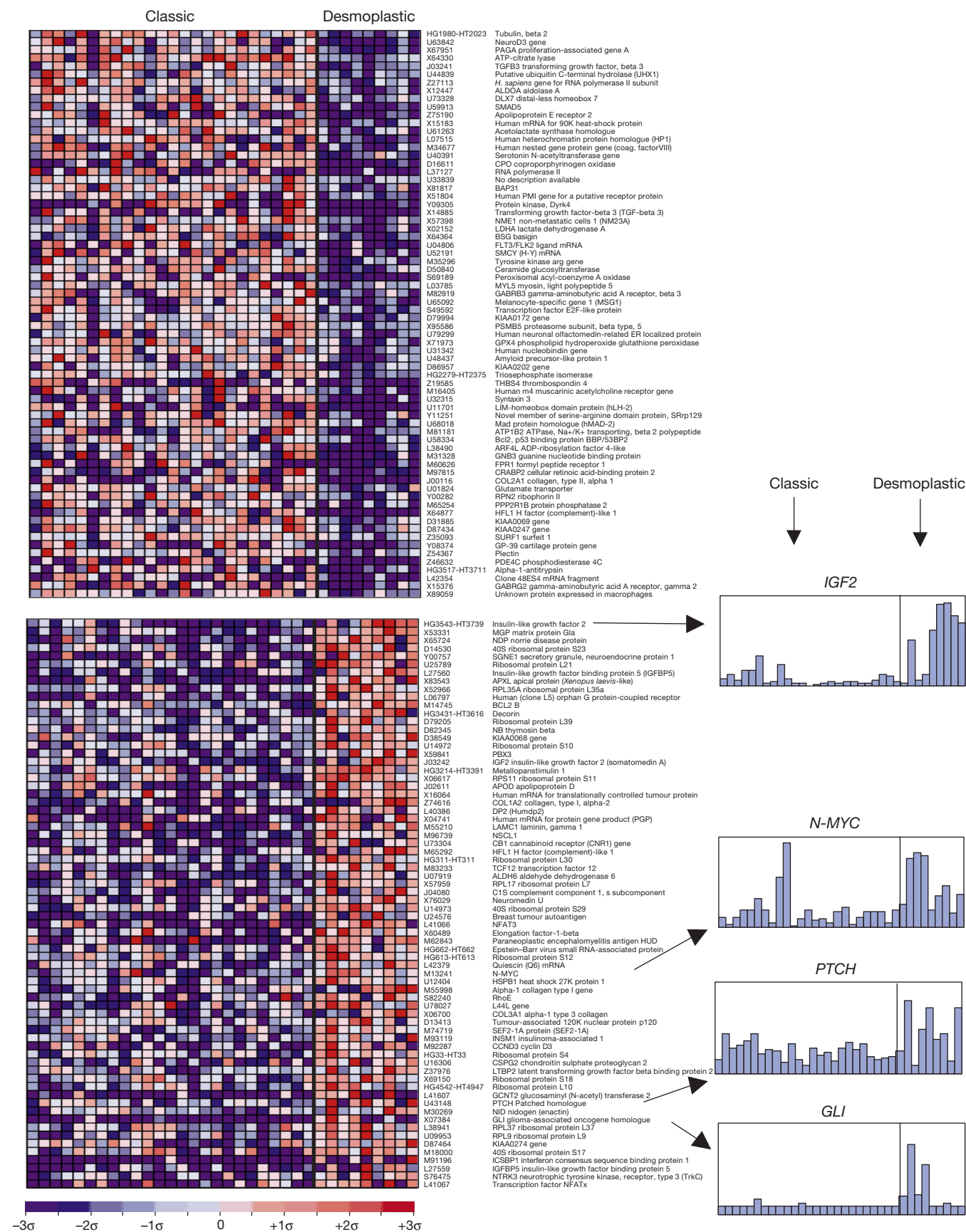


Figure 2 Differential expression of genes in classic versus desmoplastic medulloblastomas. Genes were ranked by the signal-to-noise metric according to their correlation with the classic versus desmoplastic distinction. Genes shown are those more highly correlated with the distinction than 99% of permutations of the class labels ($P < 0.01$; see

Supplementary Information III). GenBank accession numbers and gene descriptions are shown. Genes regulated by SHH are shown at the right. Normalized level of expression of selected SHH-regulated genes is shown at right. Each bar represents a different tumour.

(AT/RTs) and malignant gliomas. Previously unrecognized evidence supporting the derivation of medulloblastomas from cerebellar granule cells through activation of the Sonic Hedgehog (SHH) pathway was also revealed. We show further that the clinical outcome of children with medulloblastomas is highly predictable on the basis of the gene expression profiles of their tumours at diagnosis.

We first addressed the problem of distinguishing different embryonal CNS tumours from each other. This is important as the classification of these tumours based on histopathological appearance is debated (Fig. 1A). There are two hypotheses regarding the classification of medulloblastomas: the first is that they are part of a larger class of PNETs arising from a common cell type in the subventricular germinal matrix¹, the second is that they arise from cerebellar granule cell progenitors². To begin to generate a molecular taxonomy of CNS embryonal tumours, we analysed the gene expression profiles of 42 patient samples (data set A: 10 medulloblastomas, 5 CNS AT/RTs, 5 renal and extrarenal rhabdoid tumours, and 8 supratentorial PNETs, as well as 10 non-embryonal brain tumours (malignant glioma) and 4 normal human cerebella). RNA extracted from frozen specimens was analysed with oligonucleotide microarrays containing probes for 6,817 genes. The gene expression data are available as Supplementary Information II (see also <http://www.genome.wi.mit.edu/MPR/CNS>).

To determine whether the different types of tumours could be molecularly distinguished, we used a method of data reduction called principal component analysis in which the high dimensionality of the data was reduced to three viewable dimensions representing linear combinations of variables (genes) that account for most of the variance in the original data set (Fig. 1B)⁴. Normal brain was easily separable from the brain tumours, and the different tumour types were similarly separable. Separation of tumour types was also seen using hierarchical clustering (Fig. 1D)⁵. A more appropriate strategy for distinguishing known tumour types, however, is to use supervised learning methods to identify the genes most highly correlated with the tumour type distinctions (Fig. 1C, E). Analysis of 1,000 random permutations of the data failed to yield a separation of tumour classes to the extent observed in Fig. 1C, indicating that the observed gene expression patterns could not be explained by chance (Supplementary Information III). The robustness of these markers for classification was further investigated using a weighted-voting algorithm and evaluated by cross validation testing⁶. Correct classification of the tumours could be achieved with accuracy (35 out of 42 correct classifications, $P < 10^{-10}$ compared with random classification; see Supplementary Information III).

As expected, malignant gliomas were clearly separable from medulloblastomas, reflecting the derivation of gliomas from cells of non-neuronal origin. Consistent with this, the gliomas expressed

genes typical of the astrocytic and oligodendrocytic lineage (*PEA15*, *SOX2*, *PMP2*, *Olig-2*, TrkB kinase-negative splice variant, *S100*, *GFAP*), genes related to metabolism (fructose 2,6-bisphosphatase, glutamate dehydrogenase), and genes involved in cell differentiation (*ID2*, *GDF1*, *TYK2*; Fig. 1E and Supplementary Information III). The medulloblastomas form a cluster that is also separate from the PNETs (Fig. 1C), supporting the hypothesis that these two classes of embryonal tumours are molecularly distinct. Among the genes most highly correlated with the medulloblastoma class were *ZIC* and *NSCL1*, encoding transcription factors that are specific for cerebellar granule cells (Fig. 1E)^{7,8}. This result suggests that medulloblastomas, but not PNETs, arise from cerebellar granule cells, or alternatively, have activated the transcriptional programme of cerebellar granule cells.

We next analysed the AT/RT tumours, which have only recently been distinguished from medulloblastoma⁹. Accurate identification of AT/RT is particularly important because patients with these tumours have an extremely poor prognosis. AT/RT tumours arise in the CNS or in other organs such as the kidney, where they are referred to as rhabdoid tumours. Most tumours harbour *hSNF5/INI1* mutations, but it is unknown whether AT/RTs arising in different anatomical locations are molecularly distinct^{9–11}. As shown in Fig. 1C, the AT/RTs and rhabdoid tumours were easily distinguishable from the other tumour types in the study. Of note, the CNS AT/RTs and abdominal rhabdoid tumours were molecularly similar despite having arisen in different anatomical locations. This finding supports the idea that they arise from a similar cell of origin. Alternatively, a common mechanism of transformation may yield similar transcriptional programmes in cells of distinct origin. Markers of the distinction between AT/RT and rhabdoid tumours include genes specifically expressed during myogenesis, including skeletal β -tropomyosin, neutral calponin, *NEAT3*, and myosin regulatory light chain (Fig. 1E and Supplementary Information III). This finding is consistent with the hypothesis that the tumours have a mesenchymal origin.

We next focused on molecular heterogeneity within a single tumour type, medulloblastoma. The principal histological subclass of medulloblastoma is desmoplastic medulloblastoma, although its diagnosis is highly subjective (Fig. 1A). Desmoplastic medulloblastoma is of interest because it is seen with high frequency in patients with Gorlin's syndrome, a rare autosomal dominant disorder resulting from mutation of the SHH receptor *PTCH*^{12,13}. It is unclear whether dysregulation of the SHH pathway, known to be mitogenic for cerebellar granule cells, is also involved in the pathogenesis of sporadic desmoplastic medulloblastoma^{14–18}.

To determine whether desmoplastic and classic medulloblastoma are distinguishable by gene expression, we analysed 34 medulloblastoma samples (data set B) whose histology was scored using World Health Organization (WHO) criteria¹⁹. As shown in Fig. 2, a sharp and statistically significant gene expression signature of desmoplastic histology was evident, and this signature was sufficient for correct classification of 33 out of 34 tumours ($P = 8.6 \times 10^{-7}$ compared with random classification; see Supplementary Information III). Notably, among the genes most highly correlated with desmoplastic medulloblastoma was *PTCH* (itself a transcriptional target of SHH), as well as two other SHH downstream targets: *GLI2*²⁰ and *N-MYC* (A. Kenney and D. Rowitch, personal communication). Furthermore, insulin-like growth factor II expression was correlated with desmoplastic histology, and its expression is essential for SHH-mediated tumorigenesis in mice²¹. Taken together, the transcriptional profiling indicates that sporadic desmoplastic medulloblastomas, like tumours associated with Gorlin's syndrome, are characterized by activation of the SHH signalling pathway, further supporting the proposal that SHH dysregulation may be important in the pathogenesis of medulloblastoma.

A clinical challenge concerning medulloblastoma is the highly variable response of patients to therapy. Whereas some patients are

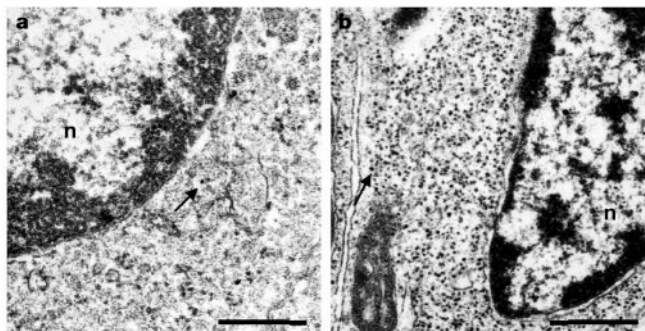


Figure 3 Representative electron micrographs showing medulloblastomas with low ribosome (a) and high ribosome (b) content. Each panel shows a portion of a single cell with a portion of the nucleus (n) (arrows designate ribosomes). Scale bars, 0.5 μ m.

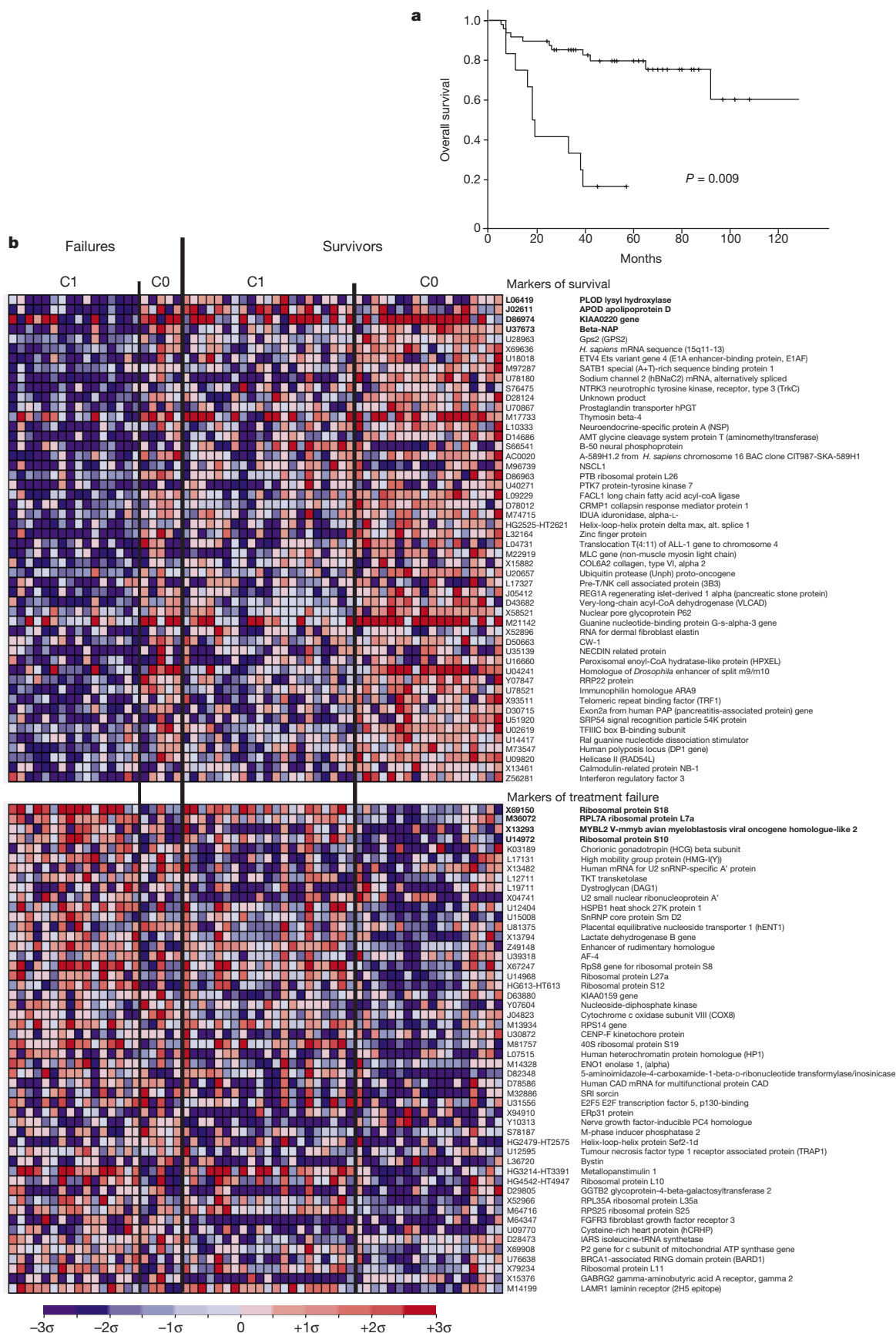


Figure 4 Predicting medulloblastoma outcome by gene expression profiling. **a**, Kaplan–Meier overall survival curves for patients predicted to survive and patients predicted to be treatment failures using an 8-gene k -NN model. Vertical hash marks indicate time of censorship. **b**, Fifty genes most highly associated with favourable outcome (top panel) or with treatment failure (bottom panel) according to the signal-to-noise metric. Samples are

further sorted according to their membership in the two unsupervised SOM-derived clusters (C0, C1). Class C1 tumours are notable for their high ribosomal content. The 8 genes most frequently used by the k -NN outcome predictor are indicated in bold. The colour scheme is the same as Fig. 1E.

cured by chemotherapy and radiation, others have progressive disease³. Currently, the only prognostic factor used in clinical practice is tumour staging; a reflection of postoperative tumour size and the presence of metastases. Unfortunately, staging-based prognostication is imperfect in that many patients with low-stage disease still succumb to their disease. There are currently no molecular markers of outcome used in clinical practice for any brain tumour. High levels of expression of the neurotrophin-3 receptor (TRKC), however, have been reported to correlate with a favourable medulloblastoma outcome^{22,23}, suggesting a molecular basis for the variability of medulloblastoma outcome. Molecular correlates of medulloblastoma metastasis have also been recently reported²⁴.

To explore the heterogeneity in response to treatment of medulloblastomas, we expanded our analysis to include 60 similarly treated patients from whom biopsies were obtained before receiving treatment, and for whom clinical follow-up was available (data set C). We first investigated whether clustering methods would identify biologically distinct subsets of the tumours. The tumours were clustered into two groups using self-organizing maps (SOM); an unsupervised algorithm that groups samples into a predetermined number of clusters on the basis of their gene expression patterns^{6,25}. The genes most highly correlated with the SOM clusters were primarily ribosomal protein-encoding genes (Supplementary Information III), suggesting differences in ribosome biogenesis. Blinded electron microscopic examination of 9 samples by 3 observers confirmed that tumours falling into the cluster characterized by high expression of ribosomal protein genes contained higher numbers of ribosomes ($P = 0.03$, Fisher's exact test; Fig. 3). We next investigated whether the SOM-derived clusters were correlated with patient survival. No statistically significant difference in the proportion of survivors compared with treatment failures in each cluster was observed (Fisher's exact test, $P = 0.1$; see Supplementary Information III). Because unsupervised methods are generally not the most appropriate analytical approach to predicting known distinctions such as outcome, we developed a supervised learning outcome predictor based on gene expression in which the classifier 'learns' the distinction between patients who are alive after treatment ('survivors') compared with those who succumbed to their disease ('failures', minimum follow-up of 24 months for surviving patients, overall median 41.5 months).

We used a k -nearest neighbours (k -NN) algorithm²⁶ that computes the distance of a test sample to each of the training set samples, each of which has an associated class (in this case, survivor or failure), and then predicts the class of the test sample to be that of the majority of the k -closest samples. The k -NN classifier was evaluated by cross-validation, whereby one sample is randomly withheld, a model is trained on the remaining samples, and the model is then used to predict the class of the withheld sample. The process is repeated until all of the samples are tested.

Outcome predictions based on gene expression were statistically significant for k -NN models ranging from 2 to 21 genes, with optimal predictions made by an 8-gene model that made only 13/60 classification errors (Fisher's exact test, $P = 0.0002$). Shown most clearly by a Kaplan–Meier survival analysis in Fig. 4a, patients that were predicted to be survivors had a 5 year overall survival of 80% compared with 17% for patients predicted to have a poor outcome ($P = 0.000003$, log rank test). A more conservative method of assessing statistical significance is to attempt to optimize classifiers of random permutations of the survivor/failure class labels. We performed 1,000 such permutations and found only 9 for which prediction accuracy matched or exceeded our observed result (Supplementary Information III), indicating that the result is unlikely to be achieved by chance ($P = 0.009$). We subsequently tested several other classification algorithms including weighted voting^{6,27}, support vector machines^{28,29} and IBM SPLASH³⁰, all of which performed with similarly high accuracy (Supplementary Information I and III).

We explored further the clinical value of the predictor by considering existing prognostic factors for medulloblastoma outcome. Patients with localized disease (M0) had a more favourable outcome compared with patients with involvement of the cerebrospinal fluid or with distant metastases (M+) ($P = 0.03$ comparing M0 with M+ by Kaplan–Meier analysis), although not all M0 patients survived. When our outcome predictor was applied only to the 42 M0 patients, the prediction of outcome remained significant ($P = 0.002$), indicating that the expression-based predictor substantially improved staging-based prognostication. Similarly, prediction based on TRKC expression was imperfect in this series in that not all patients in the unfavourable (TRKC-low) category died. When our gene expression-based predictor was applied to the 33 TRKC-low patients, the surviving patients could be significantly separated from those who succumbed to their disease ($P = 0.01$, Supplementary Information III). Of note, not all patients in this study received identical therapy. However, restricting our analysis to the 35 patients that received surgery, vincristine, cisplatin and cyclophosphamide, the predictor continued to yield a significant Kaplan–Meier survival distinction ($P = 0.0012$). Taken together, these results demonstrate that the outcome predictor based on gene expression exceeds other approaches to prognosis determination.

A number of genes not previously associated with clinical outcome were identified (Fig. 4b). Those correlated with favourable outcome included many genes characteristic of cerebellar differentiation (vesicle coat protein β -NAP, NSCL1, TRKC, sodium channels), and genes encoding extracellular matrix proteins (PLOD lysyl hydroxylase, collagen type V α 1, elastin). As expected, TRKC expression was correlated with a favourable outcome, consistent with previous reports of this association^{22,23}. In contrast, genes related to cerebellar differentiation were underexpressed in poor prognosis tumours, which were dominated by the expression of genes related to cell proliferation and metabolism (MYBL2, enolase 1, LDH, HMG1(Y), cytochrome C oxidase) and multidrug resistance (sorcin). Genes correlated with poor outcome included a number of the ribosomal protein-encoding genes identified by the SOM clustering experiments (Fig. 4b). This indicates that whereas this ribosomal signature is correlated with poor outcome, optimal outcome prediction requires not only these genes, but also genes uncorrelated with a favourable outcome that were not identified by the unsupervised clustering analysis.

The routine clinical implementation of genomics-based outcome predictors must await confirmation in independent data sets, and the models may need to be modified as treatment regimens evolve. For patients predicted to have a favourable outcome, efforts to minimize toxicity of therapy might be indicated, whereas for those predicted not to respond to standard therapy, earlier treatment with experimental regimens might be considered. This work illustrates how genomic technologies have the potential to advance treatment planning beyond the empiric, towards a more molecularly defined, individualized approach to medicine. □

Methods

Patient samples

Patients included 60 children with medulloblastomas, 10 young adults with malignant gliomas (WHO grades III and IV), 5 children with AT/RTs, 5 with renal/extrarenal rhabdoid tumours, and 8 children with supratentorial PNETs (see Supplementary Information I). Medulloblastoma patients were treated with craniospinal irradiation to 2,400–3,600 centigray (cGy) with a tumour dose of 5,300–7,200 cGy. All patients with medulloblastoma were treated with chemotherapy consisting of cisplatin and vincristine, plus combinations of carboplatin, etoposide, cyclophosphamide or lumustine (1-(2-chloroethyl)-3-cyclohexyl-1-nitrosourea, CCNU) (details in Supplementary Information II). Samples were snap frozen in liquid nitrogen and stored at -80°C . Studies were done with approval of the Committee for Clinical Investigation of Boston Children's Hospital. The data were organized into three sets: data set A (42 samples containing 10 medulloblastomas, 10 malignant gliomas, 10 AT/RTs, 8 PNETs and 4 normal cerebella); data set B (34 samples containing 9 desmoplastic medulloblastomas and 25 classic medulloblastomas); and data set C (60 samples containing 39 medulloblastoma survivors and 21 treatment failures). The clinical attributes of each of the patients in the study are available

in Supplementary Information II. Tissues were homogenized in guanidinium isothiocyanate and RNA was isolated by centrifugation over a CsCl gradient. RNA integrity was assessed either by northern blotting or by gel electrophoresis. Ten–twelve micrograms total RNA was used to generate biotinylated antisense RNAs, which were hybridized overnight to HuGeneFL arrays containing 5,920 known genes and 897 expressed sequence tags, as described previously⁶. Arrays were scanned on Affymetrix scanners and the expression value for each gene was calculated using GENECHIP software (Affymetrix, Santa Clara, California). Minor differences in microarray intensity were corrected using a linear scaling method as detailed in Supplementary Information I. Scans were rejected if the scaling factor exceeded 3, fewer than 1,000 genes received ‘present’ calls, or microarray artefacts were visible.

Preprocessing and clustering

The gene expression data were subjected to a variation filter that excluded genes showing minimal variation across the samples being analysed, as detailed in Supplementary Information I.

The data were first normalized by standardizing each column (sample) to mean 0 and variance 1. SOMs were performed using our GeneCluster clustering package (<http://www.genome.wi.mit.edu/MPR/software>). Hierarchical clustering was performed using Cluster and TreeView software⁵. Principal component analysis (PCA) was performed by computing and then plotting the three principal components using the S-Plus statistical software package (www.insightful.com/products/desktop.asp) using default settings.

Supervised learning

Genes correlated with particular class distinctions (for example, classic versus desmoplastic medulloblastoma) were identified by sorting all of the genes on the array according to the signal-to-noise statistic $(\mu_0 - \mu_1)/(\sigma_0 + \sigma_1)$, where μ and σ represent the median and standard deviation of expression, respectively, for each class. Similar results were obtained using a standard *t*-statistic as the metric $((\mu_0 - \mu_1)/\sqrt{(\sigma_0^2/N_0 + \sigma_1^2/N_1)})$, where *N* represents the number of samples in each class (see Supplementary Information). Permutation of the column (sample) labels was performed to compare these correlations to what would be expected by chance in 99% of the permutations. For classification, we developed a modification of the *k*-NN algorithm²⁶ that predicts the class of a new data point by calculating the euclidean distance (*d*) of the new sample to the *k* nearest samples (for these experiments we set *k* = 5) in the training set using normalized gene expression data, and selecting the class to be that of most of the *k* samples. The weight given to each neighbour was $1/d$. The *k*-NN models were evaluated by 60-fold ‘leave-one-out’ cross-validation, whereby a training set of 59 samples was used to predict the class of a randomly withheld sample, and the cumulative error rate was recorded. We tested models with variable numbers of genes (1–200, selected according to their correlation with the survivor versus treatment failure distinction in the training set) in this manner. An 8-gene *k*-NN outcome prediction model yielded the lowest error rate, and was therefore used to generate Kaplan–Meier survival plots using S-Plus. Predictors using metastatic staging or *TRKC* expression were constructed by finding the decision boundary half way between the classes, $(\mu_{\text{class0}} + \mu_{\text{class1}})/2$ using either the staging values 0 versus 1, 2, 3, 4 or the continuous *TRKC* microarray gene expression levels, and then predicting the unknown sample according to its location with respect to that boundary.

Received 18 July; accepted 13 November 2001.

- Rorke, L. B. The cerebellar medulloblastoma and its relationship to primitive neuroectodermal tumors. *J. Neuropathol. Exp. Neurol.* **42**, 1–15 (1983).
- Kadin, M. E., Rubenstein, L. J. & Nelson, J. S. Neonatal cerebellar medulloblastoma originating from the fetal external granular layer. *J. Neuropath. Exp. Neurol.* **29**, 583–600 (1970).
- Packer, R. J. *et al.* Treatment of children with medulloblastomas with reduced-dose craniospinal radiation therapy and adjuvant chemotherapy: a children’s cancer group study. *J. Clin. Oncol.* **17**, 2127–2136 (1999).
- Mardia, K. V., Kent, J. T. & Bibby, J. M. *Multivariate Analysis* (Academic, London, 1979).
- Eisen, M. B., Spellman, P. T., Brown, P. O. & Botstein, D. Cluster analysis and display of genome-wide expression patterns. *Proc. Natl Acad. Sci. USA* **95**, 14863–14868 (1998).
- Golub, T. R. *et al.* Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* **286**, 531–537 (1999).
- Aruga, J. *et al.* A novel zinc finger protein, Zic, is involved in neurogenesis, especially in the cell lineage of cerebellar granule cells. *J. Neurochem.* **63**, 1880–1890 (1994).
- Yokota, N. *et al.* Predominant expression of human Zic in cerebellar granule cell lineage and medulloblastoma. *Cancer Res.* **56**, 377–383 (1996).
- Rorke, L. B., Packer, R. J. & Biegel, J. A. Central nervous system atypical teratoid/rhabdoid tumors of infancy and childhood: definition of an entity. *J. Neurosurg.* **85**, 56–65 (1996).
- Biegel, J. A. *et al.* Germ-line and acquired mutations of INI1 in atypical teratoid and rhabdoid tumors. *Cancer Res.* **59**, 74–79 (1999).
- Versteeg, I. *et al.* Truncating mutations of hSNF5/INI1 in aggressive paediatric cancer. *Nature* **394**, 203–206 (1998).
- Hahn, H. *et al.* Mutations of the human homolog of *Drosophila* patched in the nevoid basal cell carcinoma syndrome. *Cell* **85**, 841–851 (1996).
- Johnson, R. L. *et al.* Human homolog of *patched*, a candidate gene for the basal cell nevus syndrome. *Science* **272**, 1668–1671 (1996).
- Pietsch, T. *et al.* Medulloblastomas of the desmoplastic variant carry mutations of the human homologue of *Drosophila* *patched*. *Cancer Res.* **57**, 2085–2088 (1997).
- Raffel, C. *et al.* Sporadic medulloblastomas contain *PTCH* mutations. *Cancer Res.* **57**, 842–845 (1997).
- Xie, J. *et al.* Mutations of the *PATCHED* gene in several types of sporadic extracranial tumors. *Cancer Res.* **57**, 2369–2372 (1997).
- Wechsler-Reya, R. J. & Scott, M. P. Control of neuronal precursor proliferation in the cerebellum by Sonic Hedgehog. *Neuron* **22**, 103–114 (1999).

- Wetmore, C., Eberhart, D. E. & Curran, T. The normal patched allele is expressed in medulloblastomas from mice with heterozygous germ-line mutation of *patched*. *Cancer Res.* **60**, 2239–2246 (2000).
- Giangaspero, F. *et al.* in *World Health Organization Histological Classification of Tumours of the Nervous System* (eds Kleihues, P. & Cavenee, W. K.) 129–137 (International Agency for Research on Cancer, Lyon, 2000).
- Murone, M., Rosenthal, A. & deSavigne, F. J. Sonic hedgehog signaling by the patched-smoothened receptor complex. *Curr. Biol.* **28**, 76–84 (1999).
- Hahn, H. *et al.* Patched target IGF2 is indispensable for the formation of medulloblastoma and rhabdomyosarcoma. *J. Biol. Chem.* **275**, 28341–28344 (2000).
- Segal, R. A., Goumnerova, L. C., Kwon, Y. K., Stiles, C. D. & Pomeroy, S. L. Expression of the neurotrophin receptor *TRKC* is linked to a favorable outcome in medulloblastoma. *Proc. Natl Acad. Sci. USA* **91**, 12867–12871 (1994).
- Kim, J. Y. H. *et al.* Activation of neurotrophin-3 receptor *TRKC* induces apoptosis in medulloblastomas. *Cancer Res.* **59**, 711–719 (1999).
- MacDonald, T. J. *et al.* Expression profiling of medulloblastoma: PDGFRA and the RAS/MAPK pathway as therapeutic targets for metastatic disease. *Nature Genet.* **29**, 143–152 (2001).
- Tamayo, P. *et al.* Interpreting patterns of gene expression with self-organizing maps: Methods and application to hematopoietic differentiation. *Proc. Natl Acad. Sci. USA* **96**, 2907–2912 (1999).
- Dasarthy, V. B. (ed). *Nearest Neighbor (NN) Norms: NN Pattern Classification Techniques* (IEEE Computer Society Press, Los Alamitos, California, 1991).
- Slonim, D. K. *et al.* in *Proc. 4th Annu. Int. Conf. Computational Mol. Biol.* 263–272 (ACM Press, New York, 2000).
- Mukherjee, S. *et al.* Support vector machine classification of microarray data. CBCL paper 182/AI memo 1676 (Massachusetts Institute of Technology, Cambridge, Massachusetts, 1999); also at <http://www.ai.mit.edu/projects/cbcl/publications/ps/cancer.ps>.
- Brown, M. P. S. *et al.* Knowledge-based analysis of microarray gene expression data by using support vector machines. *Proc. Natl Acad. Sci. USA* **97**, 262–267 (2000).
- Califano, *et al.* in *Proc. 8th Int. Conf. Intel. Syst. Mol. Biol.* (eds Bourne, P. *et al.*) 75–85 (AAAI Press, Menlo Park, CA, 2000).

Supplementary Information accompanies the paper on Nature’s website (<http://www.nature.com>).

Acknowledgements

We thank members of the Whitehead/MIT Center for Genome Research, Program in Cancer Genomics, and J. Volpe for discussions and comments on the manuscript. This work was supported in part by Millennium Pharmaceuticals, Affymetrix and Bristol-Myers Squibb (E.S.L.); NIH grants (S.L.P. and T.C.); NIH-supported Mental Retardation Research Center (S.L.P.) and Cancer Center Support CORE (T.C.); the American Lebanese Syrian Associated Charities (ALSAC); and the Kyle Mullarkey Medulloblastoma Research Fund. We acknowledge the Cooperative Human Tissue Network and the Children’s Oncology Group for contributing tumour samples.

Competing interests statement

The authors declare competing financial interests: details accompany the paper on Nature’s website (<http://www.nature.com>).

Correspondence and requests for materials should be addressed to S.L.P. (e-mail: scott.pomeroy@tch.harvard.edu) or T.R.G. (e-mail: golub@genome.wi.mit.edu).

A calcium sensor in the sodium channel modulates cardiac excitability

Hanno L. Tan[†], Sabina Kupersmidt[‡], Rong Zhang[‡], Svetlana Stepanovic^{*}, Dan M. Roden[‡], Arthur A. M. Wilde[†], Mark E. Anderson[‡] & Jeffrey R. Balser[†]

Departments of ^{*}Anesthesiology, [§]Medicine, and [‡]Pharmacology, Vanderbilt University School of Medicine, Nashville 37232, Tennessee, USA
[†]The Experimental and Molecular Cardiology Group, Academic Medical Center, University of Amsterdam, 1105 AZ, Amsterdam, the Netherlands

Sodium channels are principal molecular determinants responsible for myocardial conduction and maintenance of the cardiac rhythm. Calcium ions (Ca²⁺) have a fundamental role in the coupling of cardiac myocyte excitation and contraction, yet mechanisms whereby intracellular Ca²⁺ may directly modulate Na channel function have yet to be identified. Here we show that calmodulin (CaM), a ubiquitous Ca²⁺-sensing protein, binds to

the carboxy-terminal 'IQ' domain¹ of the human cardiac Na channel (hH1) in a Ca^{2+} -dependent manner. This binding interaction significantly enhances slow inactivation—a channel-gating process linked to life-threatening idiopathic ventricular arrhythmias^{2,3}. Mutations targeted to the IQ domain disrupted CaM binding and eliminated Ca^{2+} /CaM-dependent slow inactivation, whereas the gating effects of Ca^{2+} /CaM were restored by intracellular application of a peptide modelled after the IQ domain. A naturally occurring mutation (A1924T) in the IQ domain altered hH1 function in a manner characteristic of the Brugada arrhythmia syndrome^{4,5}, but at the same time inhibited slow inactivation induced by Ca^{2+} /CaM, yielding a clinically benign (arrhythmia free) phenotype.

Through binding Ca^{2+} in a cooperative manner (by means of four EF-hand motifs), CaM undergoes conformational rearrangements that allow interactions with substrates participating in a wide range of adaptive responses⁶. Ca^{2+} -dependent CaM binding to the C terminus of sarcolemmal calcium channels allows Ca-dependent inactivation, which provides important feedback control over intracellular Ca^{2+} (refs 1, 7, 8). Voltage-gated Na channels also contain putative C-terminal CaM-binding domains⁹, and here we have demonstrated a specific binding interaction between the C terminus of hH1 and CaM using a yeast two-hybrid assay (see Supplementary Information).

Using a peptide antagonist of Ca^{2+} -dependent CaM binding (peptide 290–309, Fig. 1)¹⁰, we studied whole-cell hH1 currents (I_{Na}) in transfected tsA201 cells. When Na channels are depolarized,

the activation gating processes first open the channel; however, inactivation gating processes rapidly ensue to impede current flow, producing the biphasic I_{Na} transient (Fig. 1b, inset). Intracellular (patch-electrode) addition of peptide 290–309 in $1\ \mu\text{M}$ free Ca^{2+} evoked a 6 mV depolarizing shift in Na channel inactivation (Fig. 1a). The peptide had no effect on the voltage-dependence of channel activation (see legend to Fig. 1b). Consistent with previous evidence that binding of peptide 290–309 to CaM is dependent on Ca^{2+} (ref. 11), we found no inactivation voltage shift from either $1\ \mu\text{M}$ free Ca^{2+} or peptide 290–309 added alone (see legend to Fig. 1a).

Recognizing that depolarized Na channels exhibit several inactivation processes with distinct kinetic signatures, we analysed the effects of peptide 290–309 on these inactivation components individually. 'Fast' inactivation evokes the rapid decay of I_{Na} on depolarization (τ_1 , Fig. 1b, inset). Addition of 290–309 peptide plus $1\ \mu\text{M}$ free Ca^{2+} slightly increased the rate of I_{Na} decay (reducing the fitted time constant τ_1 , Fig. 1b); however, exposure to $1\ \mu\text{M}$ free Ca^{2+} alone, or even to CaM plus $1\ \mu\text{M}$ free Ca^{2+} , increased the rate of fast inactivation to a similar degree (see legend to Fig. 1b). Thus, the effects of Ca^{2+} and peptide 290–309 on the rate of fast inactivation are not specifically dependent on CaM, and are therefore mechanistically distinct from CaM-dependent effects on the overall voltage dependence of inactivation in Fig. 1a.

We therefore examined a more stable, 'slow' component of inactivation^{2,3} with intermediate kinetic properties ($\tau \approx 100$ s of milliseconds) that correspond to the period of a typical cardiac

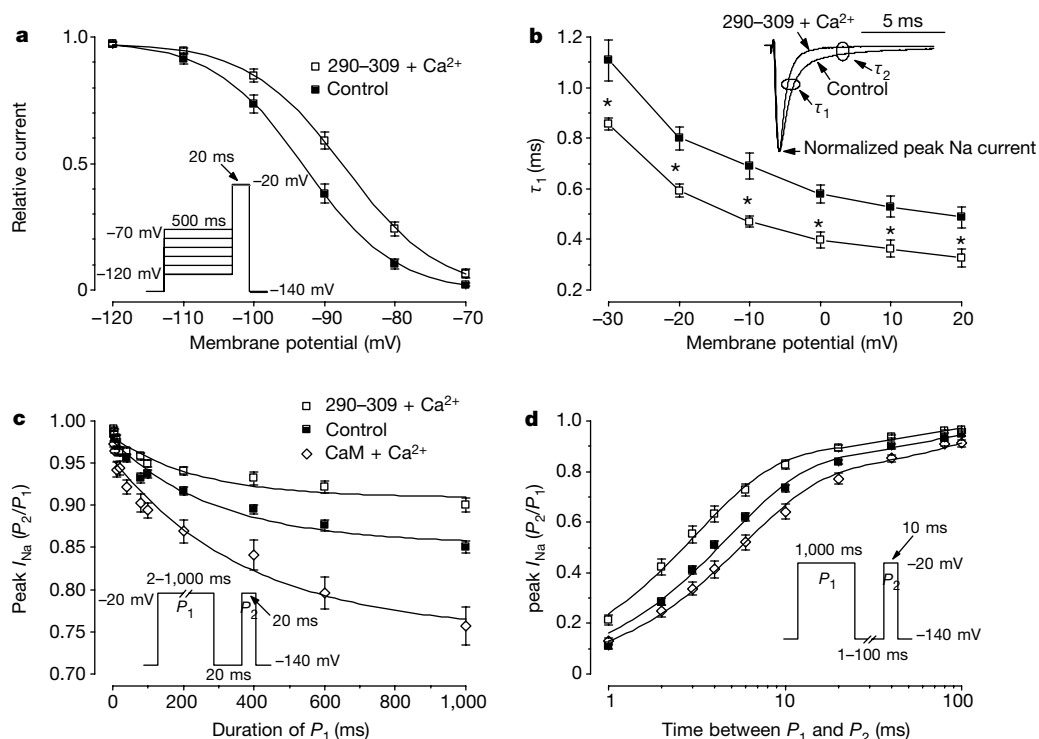


Figure 1 Inactivation gating of hH1 is sensitive to Ca^{2+} and CaM. **a**, Voltage-dependence of inactivation (see Methods). Solid lines indicate a least-squares fit of a Boltzmann function $y = [1 + \exp\{(V - V_{1/2})/k_B\}]^{-1}$ to the data. Control ($n = 13$): $V_{1/2} = -93.5 \pm 1.2$ mV, $k_B = 6.6 \pm 0.4$. Peptide 290–309 plus $1\ \mu\text{M}$ free Ca^{2+} ($n = 13$): $V_{1/2} = -87.5 \pm 1.1$ mV, $k_B = 6.6 \pm 0.4$ ($P < 0.001$ compared with control). Data not shown: the 290–309 peptide alone ($n = 5$), $V_{1/2} = -94.1 \pm 1.4$ mV, $k_B = 5.2 \pm 0.2$; $1\ \mu\text{M}$ Ca^{2+} alone ($n = 6$), $V_{1/2} = -91.4 \pm 0.8$ mV, $k_B = 4.9 \pm 0.2$. **b**, Inset, I_{Na} at -20 mV under control conditions (Methods) and with intracellular 290–309 peptide plus $1\ \mu\text{M}$ free Ca^{2+} . Both currents are normalized to the control peak. **b**, I_{Na} decays recorded at each voltage were fitted by: $y = A_1(1 - \exp[-t/\tau_1]) + A_2(1 - \exp[-t/\tau_2])$, where t is time and τ_1 and τ_2 are fitted time constants. Symbols follow the same convention as **a**. Asterisk, $P < 0.05$ compared with

control. τ_1 at $+20$ mV (plotted): control ($n = 13$), 0.50 ± 0.04 ms; peptide 290–309 plus $1\ \mu\text{M}$ free Ca^{2+} ($n = 6$), 0.32 ± 0.04 ms ($P < 0.05$ compared with control). Data not shown: $1\ \mu\text{M}$ free Ca^{2+} alone ($n = 5$), 0.33 ± 0.01 ms ($P < 0.05$ compared with control); $10\ \mu\text{M}$ CaM plus $1\ \mu\text{M}$ free Ca^{2+} ($n = 5$), 0.34 ± 0.02 ms ($P < 0.05$ compared with control). Peak I_{Na} at each depolarized membrane potential (not shown) was corrected for driving force³⁰ and fitted by a Boltzmann function (as in Fig. 1a) to assess activation: $V_{1/2}$ (mV): control, -40.3 ± 0.80 ($n = 13$); peptide 290–309 plus $1\ \mu\text{M}$ Ca^{2+} , -42.4 ± 1.34 ($n = 8$, $P = \text{NS}$ compared with control). **c**, **d**, Development and recovery from inactivation, respectively (see Methods for discussion of clamp protocols). The solid lines indicate fits to exponential functions (see Supplementary Information). The error bars in all figures refer to the standard error of the mean.

action potential, but are nonetheless distinctive from 'ultraslow' inactivation components that develop over a period of many seconds to minutes^{12–14}. Development of slow inactivation was measured using sustained depolarizations of incremental duration (Fig. 1c, protocol inset). Na channel availability ($I_{Na} (P_2/P_1)$, Fig. 1c; see Methods for definition) was increased by peptide 290–309 plus 1 μM free Ca^{2+} , but not by addition of 1 μM free Ca^{2+} alone or peptide 290–309 alone (Fig. 2, left), suggesting that slow inactivation was specifically reduced by a Ca^{2+} -dependent effect of the CaM antagonist. Although this gating effect of the CaM antagonist peptide could be indirect, we found that direct addition of intracellular (100 μM) CaM plus Ca^{2+} had the opposite effect, that is, to increase slow inactivation (Figs 1c and 2, left), suggesting that this gating process is enhanced directly by CaM.

The time constants for development and recovery from slow inactivation were examined by fitting exponential equations to the data (Fig. 1c, d, respectively). The time constant for developing slow inactivation (at -20 mV) was not significantly changed by peptide 290–309 ($\tau \approx 295$ ms, see Supplementary Information Table), although the amplitude (A) of this component was reduced. Similarly, although the overall rate of recovery from inactivation was increased by 290–309 (Fig. 1d), the time constant for recovery from slow inactivation was not significantly changed by the peptide ($\tau_2 \sim 90$ ms, see Supplementary Information Table). Instead, the relative amplitude (A_2) of the slow recovery component was reduced, reflecting partitioning of channels during the preceding (P_1) 1,000-ms depolarization step into fast-inactivated states, in preference to slow-inactivated states. Therefore, the 290–309 peptide modified steady-state partitioning into this slow, unavailable state, but did not seem to alter the rate-limiting steps controlling development or recovery from slow inactivation.

The reduction of slow inactivation by peptide 290–309 was specifically related to its interaction with CaM, because a control

peptide of similar size, but without the capability to bind CaM (see Methods), did not change slow inactivation when added with 1 μM free Ca^{2+} (Fig. 2, left). Although exogenous CaM/ Ca^{2+} increased slow inactivation, the CaM inhibitor (peptide 290–309) was sufficient to reduce slow inactivation in the absence of added CaM (Fig. 2, left), suggesting that endogenous CaM may be bound constitutively to the Na channel IQ domain, analogous to previous observations for CaM binding to Ca channels^{1,7} and Ca-dependent K channels^{15,16}. In this case, the 290–309 peptide may 'strip' endogenous CaM from the hH1 channel, and thereby reduce slow inactivation (Fig. 1c, d). Consistent with this possibility, competitive gel-shift assays using CaM, 290–309 peptide, and an IQ peptide (see Methods) bracketing the putative IQ domain in hH1 showed that peptide 290–309 binds CaM more avidly than the Na channel IQ domain (see Supplementary Information). In these binding assays, both 290–309 and IQ peptide binding to CaM required Ca^{2+} ; at the same time, in tsA201 cells slow inactivation was reduced when intracellular Ca^{2+} was minimized using BAPTA/BAPTA-AM (Fig. 2, left; see Methods). These findings suggest that the effects of slow inactivation of constitutively bound CaM are sensitive to intracellular Ca^{2+} .

We analysed further the functional interaction between Ca^{2+} , CaM and slow inactivation using hH1 mutations targeted to putative residues known to be essential to CaM binding in the IQ motif⁷ (I1908E + L1921R). In contrast to wild-type hH1, adding 290–309 peptide plus 1 μM free Ca^{2+} to tsA201 cells did not reduce slow inactivation in the hH1 mutant I1908E + L1921R (Fig. 2). Similarly, the effect of adding CaM plus free Ca^{2+} to increase slow inactivation was blunted in I1908E + L1921R (Fig. 2, middle versus left), suggesting that the mutant IQ domain was incapable of modulating Na channel gating in response to a Ca^{2+} /CaM signal. However, 10 μM CaM plus 1 μM free Ca^{2+} were sufficient to increase slow inactivation of the mutant, if added

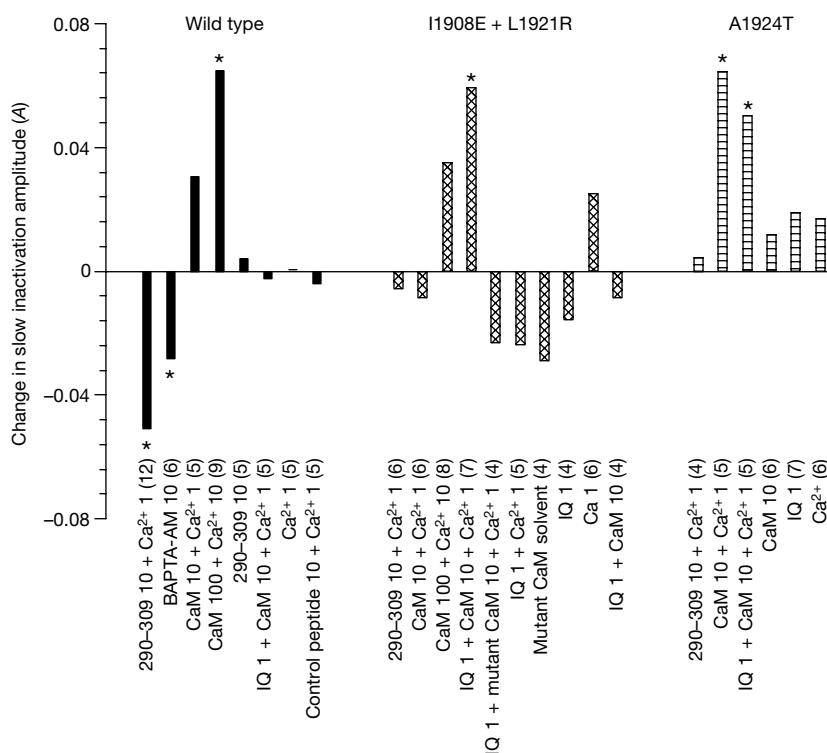


Figure 2 Conditional changes in slow inactivation in wild-type and mutant hH1 channels. Summary results are plotted for wild-type, I1908E + L1921R and A1924T. Change from the control condition within each respective group is indicated (asterisk, $P < 0.05$ compared with control), using the fitted slow inactivation amplitude term (A , see Fig. 1c

legend). For each condition, the concentration (μM) is indicated with the agent, the number of cells is indicated in parentheses, and detailed conditions are described in Methods.

with the hH1 IQ peptide (Figs 2, middle and 3a), whereas IQ peptide plus 1 μM free Ca^{2+} without CaM, or IQ peptide plus 1 μM free Ca^{2+} with a mutant CaM incapable of binding Ca^{2+} (ref. 18) had no effect on slow inactivation (Fig. 2). Furthermore, IQ peptide, Ca^{2+} , or CaM alone did not affect slow inactivation gating of the mutant (Fig. 2), and no combination of these reagents altered the rate of I1908E + L1921R fast current decay (not shown), further indicating that the effects of 290–309 peptide on fast inactivation (Fig. 1b) were independent of CaM.

The finding that exogenous IQ peptide could restore Ca^{2+} /CaM-dependent slow inactivation in the I1908E + L1921R mutant (Figs 2 and 3a) indicates that this mutation does not result in structural effects that ablate gating function. Hence, the insensitivity of the mutant Na channel may result from a direct mutational disruption of the Ca^{2+} /CaM binding site in the C-terminal IQ domain of the Na channel. Gel-shift assays (Fig. 3b) performed with CaM and the wild-type Na channel IQ peptide revealed a Ca^{2+} -dependent mobility shift, as did the homologous IQ peptide derived from the C terminus of the L-type Ca channel (see Methods), whereas a I1908E + L1921R mutant-mimetic IQ peptide did not reveal a shift, suggesting that CaM binding to the C-terminal peptide is inhibited by mutational disruption of the IQ motif.

Altered slow inactivation was recently proposed as a mechanism for arrhythmia risk in families with *SCN5A* mutations^{2,3} and a rare form of idiopathic ventricular fibrillation (Brugada syndrome)^{4,19}. Mutations in proteins regulating intracellular Ca^{2+} homeostasis

(RYR2) have also been implicated in inherited cardiomyopathic arrhythmia syndromes²⁰. To illustrate a potential pathophysiological function for the Na channel with respect to Ca^{2+} /CaM, we considered a naturally occurring *SCN5A* mutation in the IQ domain (A1924T)^{21,22}. The affected patient has no living family members who carry the mutation, but for over 40 years exhibited clear-cut Brugada syndrome electrocardiogram (ECG) characteristics (right precordial ST segment elevation and right bundle branch block) without evidence of structural heart disease. A surprising characteristic of this particular case was the absence of cardiac arrhythmias, even during the period surrounding an acute anteroseptal myocardial infarction (age 74)²².

Voltage-clamp studies of A1924T mutant hH1 channels showed that slow inactivation was reduced in baseline conditions (nominally Ca^{2+} free, Fig. 4a) and could not be reduced further by 290–309 plus 1 μM free Ca^{2+} (Fig. 2). Unlike the hH1 mutant I1908E + L1921R, CaM plus 1 μM free Ca^{2+} increased the slow inactivation of the A1924T mutant without addition of an IQ peptide (Fig. 2), and yielded approximately the same slow inactivation as the wild-type channel in baseline conditions (Fig. 4a). The IQ peptide alone had no effect on slow inactivation of the A1924T mutant, and adding the IQ peptide with CaM in 1 μM free Ca^{2+} increased slow inactivation to the same extent as CaM plus Ca^{2+} alone (Fig. 2). Binding assays (Fig. 3b) and competition assays (not shown) indicate that the affinity of an A1924T peptide for CaM is similar to the wild-type IQ peptide. Hence, the A1924T C terminus is capable of binding CaM, but is structurally modified so that the extent of slow inactivation gating in baseline conditions is reduced; hence, the net slow

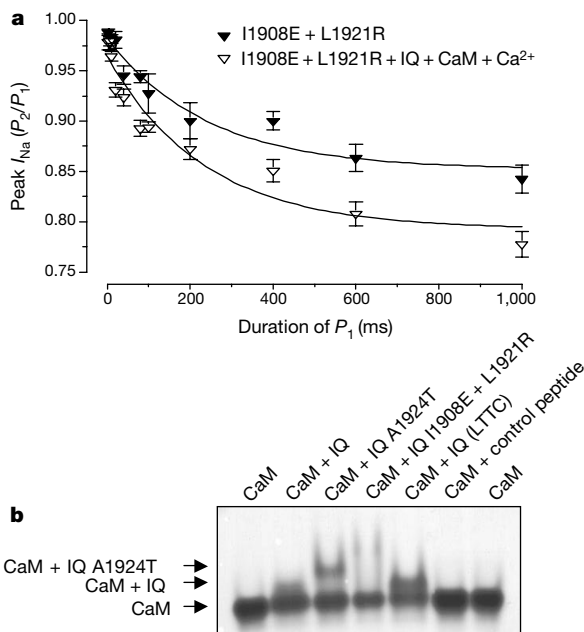


Figure 3 A mutation in the IQ domain alters Ca^{2+} /CaM sensitivity. **a**, Development of slow inactivation of I1908E + L1921R at -20 mV (protocol as in Fig. 1c). Control (filled triangles, $n = 6$), $A = 0.14 \pm 0.01$, $\tau = 288 \pm 71$ ms; 1 μM IQ peptide plus 10 μM CaM plus 1 μM Ca^{2+} (open triangles, $n = 7$), $A = 0.20 \pm 0.02$ ($P < 0.05$ compared with control), $\tau = 318 \pm 119$ ms ($P = \text{NS}$ compared with control). **b**, A gel-shift binding assay showing Ca^{2+} /CaM-dependent binding to cardiac Na⁺ channel IQ domain peptides (see Methods; no shifts were present in 1 mM EGTA with no added Ca^{2+}). All lanes contain calmodulin (CaM) and 50 μM CaCl_2 , and lanes 2–6 (left to right) also contain mimetic peptides at a fivefold molar excess, modelled after the Na⁺ channel IQ domain (IQ, lane 2), an IQ domain mutant causing an ECG Brugada phenotype (A1924T, lane 3), a double mutant in the IQ motif predicted to reduce CaM binding (I1908E + L1921R, lane 4), the cardiac L-type Ca^{2+} channel IQ domain that is known to bind CaM¹ (IQ (LTCC), lane 5), and an inactive control peptide (lane 6). Arrows indicate shifts in CaM migration in the non-denaturing gel due to CaM-peptide binding. IQ, A1924T and IQ (LTCC) all bind CaM whereas the mutant I1908E + L1921R and the control peptide both fail to alter CaM migration in the gel.

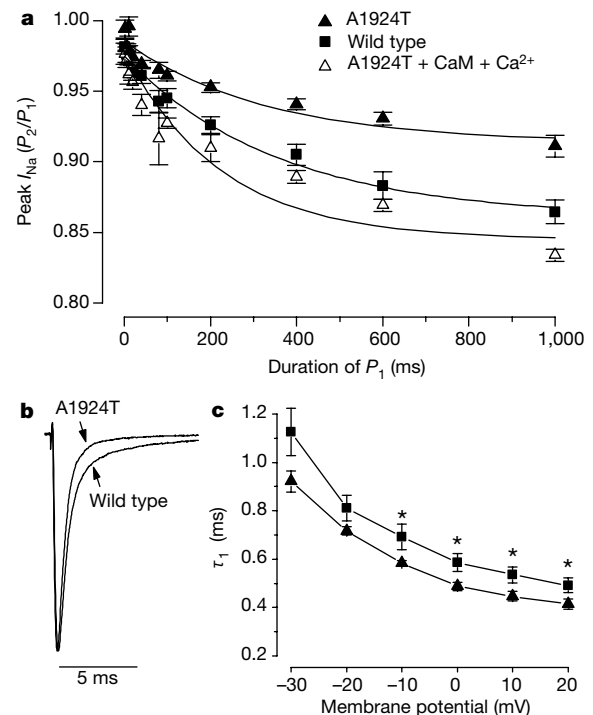


Figure 4 Inactivation gating properties of the A1924T mutation in the IQ domain. **a**, Development of slow inactivation (wild-type values re-plotted from Fig. 1c). A1924T control ($n = 8$), $A = 0.08 \pm 0.01$ ($P < 0.001$ compared with wild type), $\tau = 324 \pm 76$ ms ($P = \text{NS}$ compared with wild type); A1924T plus 10 μM CaM plus 1 μM Ca^{2+} ($n = 5$), $A = 0.14 \pm 0.01$ ($P < 0.001$ compared with A1924T control), $\tau = 426 \pm 66$ ms ($P = \text{NS}$ compared with A1924T control). Slow inactivation of wild type in control (nominally Ca^{2+} free) conditions nearly superimposes with A1924T with added CaM and Ca^{2+} . **b**, I_{Na} transients from a wild-type cell and an A1924T cell, normalized to illustrate the differences in decay rate. **c**, τ_1 is plotted (as per Fig. 1b) for wild type (squares, $n = 10$) and A1924T (triangles, $n = 4$). Asterisk, $P < 0.05$ compared with control.

inactivation in response to Ca^{2+} /CaM is also reduced.

Gating mutations associated with Brugada syndrome generally reduce Na channel availability²³, and the A1924T mutation increased the rate of fast inactivation (Fig. 4b, c) in a manner consistent with loss-of-function gating changes seen in other Brugada syndrome mutations⁵. At the same time, this C-terminal mutation attenuates Ca^{2+} /CaM-induced slow inactivation, and may therefore limit the propensity for transient changes in cytosolic Ca^{2+} or CaM to trigger cardiac arrhythmias^{24–27}. The proposal that these gating effects suppress the arrhythmia phenotype in the Brugada patient (see above) is speculative, and other mechanisms may well have an equally protective role. Nonetheless, these studies clarify that calmodulin, an ubiquitous Ca^{2+} -sensing protein, binds to the C-terminal IQ domain of the human Na^+ channel in a Ca^{2+} -dependent manner, and that this binding interaction enhances significantly a slow inactivation gating process implicated in cardiac arrhythmias^{2,3}. Moreover, these studies provide the first link, to our knowledge, between intracellular Ca^{2+} and Na channel gating function, and offer a rationale for future studies examining how this functional coupling may critically influence cellular excitability in common, acquired forms of disease involving Ca^{2+} dysregulation and disorders of cardiac rhythm. □

Methods

Mutagenesis and electrophysiology

Site-directed mutagenesis was performed on the human cardiac Na channel (hH1) α -subunit complementary DNA, cloned into the green fluorescent protein (GFP)-containing bicistronic expression vector pCGI as described previously²⁷. Cultured cells (tsA201) were transiently transfected with either wild-type or mutant α -subunit cDNA, and GFP-expressing cells were selected for electrophysiological analysis 1–2 days later. We recorded whole-cell I_{Na} at room temperature (22 °C). Voltage-dependence of inactivation was assessed as the relative current over a range of conditioning voltages (Fig. 1a). This was calculated by normalizing peak I_{Na} elicited by test depolarizations to ~ 20 mV, to the value obtained after holding at the most negative (~ 120 mV) conditioning voltage (protocol inset, Fig. 1a). Development of slow inactivation was assessed using a two-pulse voltage-clamp protocol (inset, Fig. 1c). The reduction in I_{Na} magnitude during the second test pulse (P_2), relative to that recorded during the first pulse (P_1) as the P_1 pulse is lengthened reflects occupancy of a slow inactivated state. The 20-ms recovery interval between P_1 and P_2 allows full recovery from fast inactivation, but also allows partial (~ 10 – 15%) recovery from slow inactivation, hence the magnitude of slow inactivation estimated by the analysis is a lower limit. Recovery from inactivation was assessed using an analogous protocol (Fig. 1d, inset), where the recovery interval between P_1 and P_2 was varied, and the P_1 pulse duration was fixed (1,000 ms). Other clamp protocols are described with each figure.

The bath solution contained (in mM): NaCl 145, KCl 4, CaCl_2 1.8, MgCl_2 1, HEPES 10 (pH 7.35 with NaOH). The control pipette solution was nominally Ca^{2+} free, and contained (in mM): NaF 10, CsF 110, CsCl 20, EGTA 10, and HEPES 10 (pH 7.3 with CsOH). In experiments where intracellular free Ca^{2+} was 1 μM or 10 μM , EGTA was reduced to 1 mM and CaCl_2 was added (0.90 mM and 1.0 mM, respectively)²⁸. For experiments with minimized intracellular Ca^{2+} , cells were incubated at 37 °C for 3 h in Ca^{2+} -free culture medium to which 10 μM membrane-permeable BAPTA-AM was added to chelate endogenous Ca^{2+} . Patch clamping was then performed with 20 mM BAPTA in the pipette (EGTA omitted, CsF reduced to 100 mM to maintain osmolality) and bath CaCl_2 at 1 mM.

All compounds added to the pipette solution were diluted from stock solutions dissolved in water, with the exception of mutant CaM (a gift from S. H.), which was dissolved in 50 mM MOPS plus 0.02% $\text{Na}_2\text{S}_2\text{O}_8$ (vehicle did not evoke gating effects, see Fig. 2, middle). A 50 mM stock solution of BAPTA-AM was dissolved in DMSO and diluted 1:5,000 to yield a 10 μM final concentration. We added wild-type CaM to the pipette solution to give final concentrations of 10 or 100 μM . The 290–309 CaM antagonist peptide (LKKFNARRKLGAILTIMLA; IC₅₀ 52 nM, Macromolecular Resources)¹⁰ and an inactive control peptide modelled after a non-CaM binding region of CaM kinase II (KKALHAQERVDCL)²⁹ were added to the patch electrode at a final concentration of 10 μM . (IC₅₀ is the binding affinity, the inhibitory concentration yielding 50% effect.) The 290–309 peptide numbering corresponds to the pseudosubstrate site of CaM kinase II which binds CaM with high affinity in a Ca^{2+} -dependent manner¹⁰. An IQ peptide, modelled from a region of the hH1 C terminus that exhibits homology to the L-type Ca channel (LTCC) IQ domain¹ (shown below), was added to the pipette solution at a 1 μM concentration. Mutant IQ peptides, with amino acids modified to mimic the Na channel mutants A1924T and I1908E + L1921R, were also added to the pipette solution at 1 μM (mutated amino acids are underlined).

	1908	1921	1924
Qpeptide(hH1) :	IQRARRRHLQRSLKHASFL		
LTCCIQdomain :	IQRAYRRYLKQKVKKV		

Binding assays

For gel-shift binding assays, CaM (15–30 μM) was incubated overnight with a fivefold molar excess of peptide at 4 °C in 50 mM Tris-HCl (pH 7.5) containing either 50 μM CaCl_2 or 1 mM EGTA (40 μl total volume). For competition assays, peptides were added in the molar ratios indicated (see Supplementary Information). After adding 5 μl 50% glycerol and 0.03% bromophenol blue, samples were analysed by non-denaturing polyacrylamide gel electrophoresis in the presence of 50 μM CaCl_2 (12% acrylamide) or 1 mM EGTA (12% acrylamide), followed by Coomassie blue staining.

For yeast two-hybrid analysis (see Supplementary Information) a bait vector, the C terminus (amino acids 1773 to 1917) of hH1, was constructed by polymerase chain reaction amplification and subcloning in frame with the Gal4 DNA-binding domain into pGBDU-C2 (ref. 30). The CaM coding region was isolated from a human heart cDNA library and subcloned in frame with the yeast Gal4 activation domain into the yeast prey vector pGAD-C2 (ref. 30). The minK and fhl2 control constructs were assembled using the same vector. The hH1 bait was then transformed into yeast strain PJ69-4A (ref. 30) and grown on SD/-ura (synthetic dropout media minus uracil). The prey clones were supertransformed into bait-containing PJ69-4A and tested for growth on SD/-Ura/-Leu/-His and SD/-Ura/-Leu/-Ade medium.

Received 26 October; accepted 10 December 2001.

- Zuhlke, R. D. & Reuter, H. Ca^{2+} -sensitive inactivation of L-type Ca^{2+} channels depends on multiple cytoplasmic amino acid sequences of the $\alpha_1\text{C}$ subunit. *Proc. Natl Acad. Sci. USA* **95**, 3287–3294 (1998).
- Veldkamp, M. W. *et al.* Two distinct congenital arrhythmias evoked by a multidysfunctional Na^+ channel. *Circ. Res.* **86**, E91–E97 (2000).
- Wang, D. W., Makita, N., Kitabatake, A., Balser, J. R. & George, A. L. Enhanced sodium channel intermediate inactivation in Brugada syndrome. *Circ. Res.* **87**, E37–E43 (2000).
- Chen, Q. *et al.* Genetic basis and molecular mechanism for idiopathic ventricular fibrillation. *Nature* **392**, 293–296 (1998).
- Dumaine, R. *et al.* Ionic mechanisms responsible for the electrocardiographic phenotype of the Brugada syndrome are temperature dependent. *Circ. Res.* **85**, 803–809 (1999).
- Babu, Y. S. *et al.* Three-dimensional structure of calmodulin. *Nature* **315**, 37–40 (1985).
- Lee, A. *et al.* Ca^{2+} /calmodulin binds to and modulates P/Q-type calcium channels. *Nature* **399**, 155–159 (1999).
- DeMaria, C. D., Soong, T. W., Alseikhan, B. A., Alvania, R. S. & Yue, D. T. Calmodulin bifurcates the local Ca^{2+} signal that modulates P/Q-type Ca^{2+} channels. *Nature* **411**, 484–489 (2001).
- Mori, M. *et al.* Novel interaction of the voltage-dependent sodium channel (VDSC) with calmodulin: does VDSC acquire calmodulin-mediated Ca^{2+} -sensitivity? *Biochemistry* **39**, 1316–1323 (2000).
- Payne, M. E. *et al.* Calcium/calmodulin-dependent protein kinase II. Characterization of distinct calmodulin binding and inhibitory domains. *J. Biol. Chem.* **263**, 7190–7195 (1988).
- Hanley, R. M. *et al.* Functional analysis of a complementary DNA for the 50-kilodalton subunit of calmodulin kinase II. *Science* **237**, 293–297 (1987).
- Rudy, B. Slow inactivation of the sodium conductance in squid giant axons. Pronase resistance. *J. Physiol. (Lond.)* **238**, 1–21 (1978).
- Richmond, J. E., Featherstone, D. E., Hartmann, H. A. & Ruben, P. C. Slow inactivation in human cardiac sodium channels. *Biophys. J.* **74**, 2945–2952 (1998).
- Todt, H., Dudley, S. C. Jr, Kyle, J. W., French, R. J. & Fozzard, H. A. Ultra-slow inactivation in $\mu 1$ Na^+ channels is produced by a structural rearrangement of the outer vestibule. *Biophys. J.* **76**, 1335–1345 (1999).
- Keen, J. E. *et al.* Domains responsible for constitutive and Ca^{2+} -dependent interactions between calmodulin and small conductance Ca^{2+} -activated potassium channels. *J. Neurosci.* **19**, 8830–8838 (1999).
- Xia, X. M. *et al.* Mechanism of calcium gating in small-conductance calcium-activated potassium channels. *Nature* **395**, 503–507 (1998).
- Rhoads, A. R. & Friedberg, F. Sequence motifs for calmodulin recognition. *FASEB J.* **11**, 331–340 (1997).
- Maune, J. F., Klee, C. B. & Beckingham, K. Ca^{2+} binding and conformational change in two series of point mutations to the individual Ca^{2+} -binding sites of calmodulin. *J. Biol. Chem.* **267**, 5286–5295 (1992).
- Brugada, P. & Brugada, J. Right bundle branch block, persistent ST segment elevation and sudden cardiac death: a distinct clinical and electrocardiographic syndrome. A multicenter report. *J. Am. Coll. Cardiol.* **20**, 1391–1396 (1992).
- Tiso, N. *et al.* Identification of mutations in the cardiac ryanodine receptor gene in families affected with arrhythmogenic right ventricular cardiomyopathy type 2 (ARVD2). *Hum. Mol. Genet.* **10**, 189–194 (2001).
- Rook, M. B. *et al.* Human SCN5A gene mutations alter cardiac sodium channel kinetics and are associated with the Brugada syndrome. *Cardiovasc. Res.* **44**, 507–517 (1999).
- Wilde, A. A. M. & Duren, D. Sudden cardiac death, RBBB, and right precordial ST-segment elevation. *Circulation* **99**, 722–723 (1999).
- Balser, J. R. Sodium 'channelopathies' and sudden death: must you be so sensitive? *Circ. Res.* **85**, 872–874 (1999).
- Clusin, W. T., Buchbinder, M. & Harrison, D. C. Calcium overload, 'injury' current, and early ischaemic cardiac arrhythmias—a direct connection. *Lancet* **1**, 272–274 (1983).
- Di Diego, J. M. & Antzelevitch, C. High $[\text{Ca}^{2+}]_i$ -induced electrical heterogeneity and extrasystolic activity in isolated canine ventricular epicardium. Phase 2 reentry. *Circulation* **89**, 1839–1850 (1994).
- Aimond, E., Alvarez, J. L., Raugier, J. M., Lorente, P. & Vassort, G. Ionic basis of ventricular arrhythmias in remodeled rat heart during long-term myocardial infarction. *Cardiovasc. Res.* **42**, 402–415 (1999).
- Tan, H. L. *et al.* A sodium channel mutation causes isolated cardiac conduction disease. *Nature* **409**, 1043–1047 (2001).
- Bers, D. M., Patton, C. W. & Nuccitelli, R. A practical guide to the preparation of Ca^{2+} buffers. *Methods Cell Biol.* **40**, 3–29 (1994).
- Braun, A. P. & Schulman, H. A non-selective cation current activated via the multifunctional Ca^{2+} -calmodulin-dependent protein kinase in human epithelial cells. *J. Physiol. (Lond.)* **488**, 37–55 (1995).

30. James, P., Halladay, J. & Craig, E. A. Genomic libraries and a host strain designed for highly efficient two-hybrid selection in yeast. *Genetics* **144**, 1425–1436 (1996).

Supplementary Information accompanies the paper on Nature's website (<http://www.nature.com>).

Acknowledgements

We acknowledge C. R. Bezzina for genetic analysis of A1924T, C. A. Conrath for subcloning A1924T, and A. George for critique. H.L.T. was supported by a fellowship from the Royal Netherlands Academy of Arts and Sciences. Additional financial support was provided by the Interuniversity Cardiology Institute Netherlands project 27 (H.L.T. and A.A.M.W.), the Dutch Heart Foundation NHS (A.A.M.W.), and National Institutes of Health grants (M.E.A. and J.R.B.).

Competing interests statement

The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to J.R.B. (e-mail: jeff.balser@mcmill.vanderbilt.edu).

Translocation of lipid-linked oligosaccharides across the ER membrane requires Rft1 protein

Jonne Helenius*, Davis T. W. Ng†, Cristina L. Marolda‡, Peter Walter§, Miguel A. Valvano‡ & Markus Aebl*

* Institute of Microbiology, Swiss Federal Institute of Technology, Zürich, CH-8092 Zürich, Switzerland

† Department of Biochemistry and Molecular Biology, Pennsylvania State University, University Park, Pennsylvania 16802, USA

‡ Department of Microbiology and Immunology, University of Western Ontario, London, Ontario N6A 5C1, Canada

§ Howard Hughes Medical Institute and Department of Biochemistry and Biophysics, University of California at San Francisco, San Francisco, California 94143-0448, USA

N-linked glycosylation of proteins in eukaryotic cells follows a highly conserved pathway. The tetradecasaccharide substrate ($\text{Glc}_3\text{Man}_9\text{GlcNAc}_2$) is first assembled at the membrane of the endoplasmic reticulum (ER) as a dolichylpyrophosphate (Dol-PP)-linked intermediate, and then transferred to nascent polypeptide chains in the lumen of the ER¹. The assembly of the oligosaccharide starts on the cytoplasmic side of the ER membrane with the synthesis of a $\text{Man}_5\text{GlcNAc}_2\text{-PP-Dol}$ intermediate. This lipid-linked intermediate is then translocated across the membrane so that the oligosaccharides face the lumen of the ER, where the biosynthesis of $\text{Glc}_3\text{Man}_9\text{GlcNAc}_2\text{-PP-Dol}$ continues to completion. The fully assembled oligosaccharide is transferred to selected asparagine residues of target proteins. The transmembrane movement of lipid-linked $\text{Man}_5\text{GlcNAc}_2$ oligosaccharide is of fundamental importance in this biosynthetic pathway, and similar processes involving phospholipids and glycolipids are essential in all types of cells^{2–4}. The process is predicted to be catalysed by proteins, termed flippases, which to date have remained elusive^{2–4}. Here we provide evidence that yeast *RFT1* encodes an evolutionarily conserved protein required for the translocation of $\text{Man}_5\text{GlcNAc}_2\text{-PP-Dol}$ from the cytoplasmic to the luminal leaflet of the ER membrane.

Asymmetric lipid distribution is a fundamental characteristic of biological lipid bilayers. Lipids are the building blocks of membranes, and are involved in many processes including signalling, membrane polarity, biosynthesis and intracellular traffic, where membrane sidedness is critical. Several P-type ATPase and ABC

transporters in addition to a fatty-acid transporter have been implicated in lipid translocation at the plasma membrane; however, translocators involved in membrane biogenesis have yet to be identified^{5,6}. One of the most striking examples of lipid flipping is the translocation of the $\text{Man}_5\text{GlcNAc}_2\text{-PP-Dol}$ intermediate from the cytosolic side of the ER membrane to the lumen before the completion of the biosynthesis of $\text{Glc}_3\text{Man}_9\text{GlcNAc}_2\text{-PP-Dol}$ ^{7,8} (Fig. 1). In this case, a very large and polar oligosaccharide moiety must be translocated across the hydrophobic interior of the bilayer.

There are two lines of evidence suggesting that $\text{Man}_5\text{GlcNAc}_2\text{-PP-Dol}$ is the intermediate that is translocated across the ER membrane⁷. The first is based on monosaccharide donors. Biosynthesis of $\text{Man}_5\text{GlcNAc}_2\text{-PP-Dol}$ requires membrane-associated glycosyltransferases and cytosolic nucleotide-activated sugar substrates, UDP-GlcNAc and GDP-Man. In contrast, the subsequent transferases, starting with the *ALG3* mannosyltransferase, use Dol-P-Man and Dol-P-Glc precursors that are available in the lumen of the ER⁸. The absence of GDP-mannose within the ER lumen implies that the synthesis of $\text{Man}_5\text{GlcNAc}_2\text{-PP-Dol}$ must occur on the cytosolic side of the ER membrane⁹. The second line of evidence is based on a binding study. When the mannose-specific lectin, concanavalin A, was added to intact 'right-side-out' ER vesicles, only $\text{Man}_5\text{GlcNAc}_2\text{-PP-Dol}$ and smaller intermediates bound to it, indicating that $\text{Man}_6\text{GlcNAc}_2$ and larger intermediates were sequestered in the lumen of the ER¹⁰.

The translocation of $\text{Man}_5\text{GlcNAc}_2\text{-PP-Dol}$ has long been thought to be protein-mediated. The unassisted transmembrane movement of polyisoprenol-linked sugars in artificial liposomes is extremely slow^{11,12}. The putative flippase catalysing this reaction is expected to be encoded by an essential gene, as the absence of PP-Dol-linked oligosaccharides in the ER lumen would halt all N-linked protein glycosylation—an essential process. PP-Dol-linked oligosaccharides that are incompletely assembled on the cytoplasmic side of the ER membrane (for example, $\text{Man}_5\text{GlcNAc}_2\text{-PP-Dol}$ and $\text{Man}_4\text{GlcNAc}_2\text{-PP-Dol}$) can still be transferred to the ER lumen and from there to protein, suggesting that the flippase is not strictly specific for $\text{Man}_5\text{GlcNAc}_2\text{-PP-Dol}$ but can operate, albeit inefficiently, with relaxed substrate specificity^{13,14}.

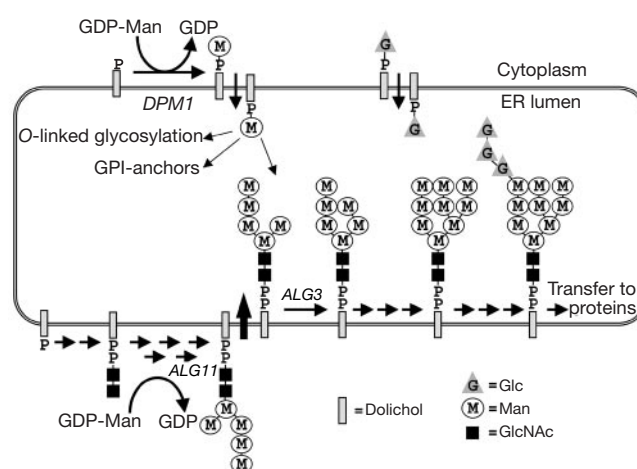


Figure 1 Biosynthesis of the N-linked oligosaccharides requires enzymatic reactions on both sides of the ER membrane. The lipid dolichylpyrophosphate serves as the membrane-bound carrier of the oligosaccharides. *ALG* gene products, including *Alg3* and *Alg11*, catalyse the stepwise transfer of single monosaccharides. On the cytoplasmic side, activated sugars are used to generate the branched oligosaccharide, $\text{Man}_5\text{GlcNAc}_2$, which is translocated into the ER. Man-P-Dol and Glc-P-Dol are synthesized on the cytoplasmic side and are also translocated into the ER where they are used in the synthesis of full-length oligosaccharide. Man-P-Dol , the product of the *DPM1* enzyme in yeast, is also required for the synthesis of GPI-anchors and O-mannosylation of proteins.

In a previous study, we identified a mutant in *Saccharomyces cerevisiae* with a screen designed to detect cellular defects that require an intact unfolded protein response¹⁵. We showed that the mutant strain was deficient in *N*-linked glycosylation¹⁵. The mutation mapped to *RFT1*, which encodes an essential protein with multiple transmembrane domains that is localized in the ER (data not shown and M. J. Cai, personal communication). Homologues of *RFT1* are found in all eukaryotic genomes sequenced to date, with the notable exception of *Plasmodium falciparum*, a eukaryotic parasite that lacks *N*-glycoproteins¹⁶.

To analyse the function of Rft1 in more detail, we constructed a haploid strain where the *GAL1-10* promoter replaced the endogenous *RFT1* promoter. Transfer of this strain to a glucose-containing medium repressed Rft1 expression and resulted in a severe reduction of growth. To assess the glycosylation defect in the mutant strain, we analysed carboxypeptidase Y (CPY), a well-characterized vacuolar protease carrying four *N*-linked oligosaccharides¹⁷. At different time points after repression of *RFT1*, CPY became increasingly hypoglycosylated, lacking up to three or even all four *N*-linked oligosaccharides (Fig. 2a). We next analysed the oligosaccharide PP-Dol precursors present in this strain at different time points after *RFT1* repression. To this end, aliquots of the same cultures were labelled with [³H]-mannose, lipid-linked oligosaccharides were isolated, and the oligosaccharide moieties were cleaved off and separated by high-performance liquid chromatography (HPLC). As shown in Fig. 2b, depletion of Rft1 resulted in a significant accumulation of the Man₅GlcNAc₂ precursor, at the expense of the more completely assembled Man₈GlcNAc₂ and Glc₃Man₉GlcNAc₂ forms.

In principle, there are several possible defects that could lead to

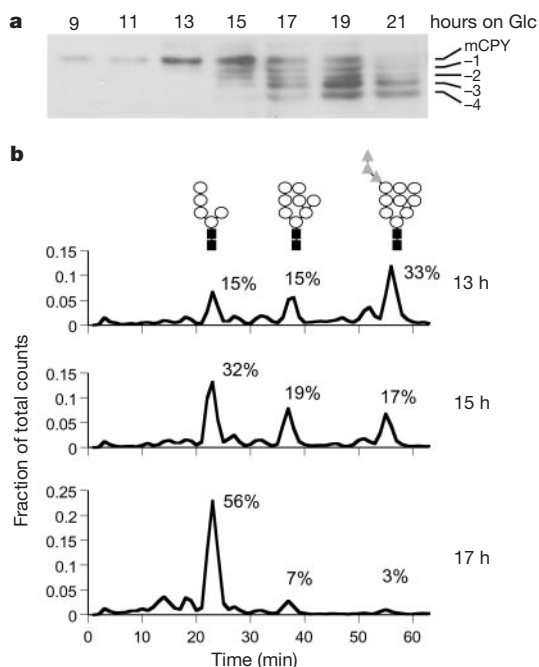


Figure 2 *RFT1* repression reduces *N*-linked glycosylation by altering the synthesis of lipid-linked oligosaccharides. The yeast strain used is YG1137, where the glucose-repressible *GAL1-10* promoter has replaced the *RFT1* promoter. **a**, Western blot analysis of CPY in YG1137 cultures grown in glucose-containing medium for the indicated times. Bands representing fully glycosylated (mCPY) and hypoglycosylated forms (lacking up to four oligosaccharides) of CPY are indicated. **b**, HPLC trace of metabolically labelled lipid-linked oligosaccharide intermediates from the same *RFT1*-repressed cells at 13, 15 and 17 h after shift to glucose-containing medium. Major HPLC peaks corresponding to Man₅GlcNAc₂, Man₈GlcNAc₂ and Glc₃Man₉GlcNAc₂ are marked. The percentage of total signal (average of two measurements) in marked peaks is indicated. Man, GlcNAc and Glc are indicated by circles, squares and triangles, respectively.

hypoglycosylation of *N*-glycoproteins with the concomitant accumulation of lipid-linked Man₅GlcNAc₂ intermediates (see Fig. 1). The effect could be explained by: (1) a deficiency in the mannosyl-transferase Alg3; (2) a decreased availability of the substrate of Alg3, Dol-P-Man; or (3) a block in translocation of the lipid-linked Man₅GlcNAc₂ intermediate across the membrane.

The first two of these possibilities are ruled out by the following observations. First, and consistent with our previous work¹⁸, Δ alg3 strains accumulate lipid-linked Man₅GlcNAc₂. Man₅GlcNAc₂ in Δ alg3 cells is transferred to proteins in the ER, resulting in CPY glycoforms that are completely resistant to endo- β -*N*-acetylglucosaminidase H (EndoH) digestion¹⁸ (Fig. 3, top panel). In contrast, Rft1 depletion resulted in hypoglycosylated CPY that remained sensitive to EndoH digestion (Fig. 3, top panel), indicating that any residual glycosylation involved the transfer of fully assembled core oligosaccharide units. Thus, *RFT1*-depleted cells did not lack *ALG3* activity.

The second observation is that the ER-luminal pool of Dol-P-Man, the substrate from which mannoses are transferred to lipid-linked oligosaccharides in the ER lumen, is not diminished. Specific leaky mutations in Dol-P-Man synthase, encoded by a single essential locus, *DPM1*, display a temperature-sensitive growth phenotype and have a very low pool of Dol-P-Man. This low level of Dol-P-Man not only affects *N*-linked but also *O*-linked protein glycosylation, as Dol-P-Man is the primary substrate in the *O*-glycosylation process in the ER of fungal cells¹⁹. As expected, *dpm1-6* mutant cells displayed a heterogeneous population of hypoglycosylated CPY molecules, indicative of a deficiency in *N*-linked protein glycosylation (Fig. 3, top panel, lanes 7 and 8). Furthermore, depletion of Dol-P-Man in the *dpm1* mutant also resulted in altered *O*-linked mannosylation, as visualized by the presence of unglycosylated or partially glycosylated chitinase in these cells (Fig. 3, bottom panel, lanes 7 and 8)²⁰. Chitinase is an extensively *O*-mannosylated protein whose mobility, as measured in an SDS-polyacrylamide gel electrophoresis (PAGE) gel, shifts from a relative molecular mass of 130,000 to 60,000 (*M*_r 130K and 60K, respectively) in the absence of *O*-mannosylation²⁰. Importantly, depletion of Rft1 had no effect on chitinase glycosylation (Fig. 3, bottom panel, lanes 3 and 4), demonstrating that there is a normal pool of Dol-P-Man in these cells. As expected, the same result was observed in the Alg3-deficient strain (Fig. 3, bottom panel, lanes 5 and 6). Depletion of Rft1 did not affect the level of the glycosylphosphatidylinositol (GPI)-anchored Gas1 protein (data not shown). Biosynthesis of the GPI-anchor is sensitive to deficient Dol-P-Man synthesis activity, because the core mannosyl residues of the GPI-

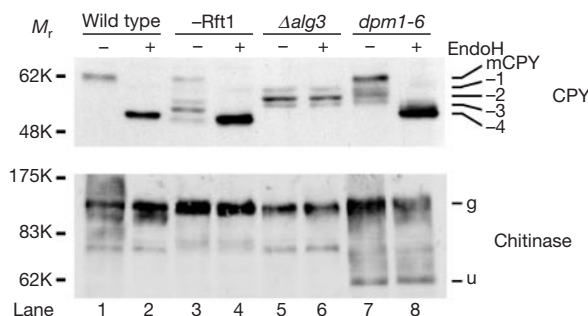


Figure 3 *RFT1* repression results in a distinct protein glycosylation defect. Western blot analysis of protein extracts from wild-type (SS328), Δ alg3 (YG248) and *dpm1-6* mutant (grown for 4 h at 37 °C) strains are shown. Protein extracts were incubated overnight in the presence (+) or absence (–) of EndoH. Blots were probed with antibodies against CPY (top panel) and chitinase (bottom panel). Mobility of fully glycosylated (mCPY) and hypoglycosylated forms of CPY are marked in the top panel, whereas fully glycosylated (g) and unglycosylated (u) forms of chitinase are specified in the bottom panel.

anchors are derived from Dol-P-Man¹⁹.

Taken together, these results point to the possibility that the phenotypes observed on Rft1 depletion result strictly from impaired Man₅GlcNAc₂-PP-Dol translocation across the ER membrane. This hypothesis is supported further by the following findings. *ALG11* encodes a GDP-Man-dependent mannosyltransferase, which is required for cytoplasmic synthesis of Man₅GlcNAc₂-PP-Dol intermediate¹⁴. $\Delta alg11$ cells grow very slowly, are temperature sensitive, and contain severely hypoglycosylated proteins. These mutant cells accumulate Man₃GlcNAc₂-PP-Dol, which is then translocated, albeit inefficiently, across the ER membrane. Once inside the ER, luminal mannosyltransferases add four additional Dol-P-Man-derived mannose residues, resulting in Man₇GlcNAc₂ oligosaccharides. Therefore, analysis of lipid-linked oligosaccharides in $\Delta alg11$ cells reveals two major peaks, representing Man₃GlcNAc₂ and Man₇GlcNAc₂ (Fig. 4a). Notably, Man₃GlcNAc₂ was shown to face the cytosol¹⁰, whereas Man₇GlcNAc₂ faces the ER lumen, suggesting that translocation into the ER lumen of the incomplete Man₃GlcNAc₂ is the rate-limiting step in the biosynthesis of Man₇GlcNAc₂ (ref. 14). We proposed that an increased amount of the postulated flippase might therefore compensate for the *ALG11* deletion-induced defect, resulting in an increased formation of Man₇GlcNAc₂-PP-Dol and protein glycosylation.

To test this hypothesis, we introduced *RFT1* on a multicopy plasmid into $\Delta alg11$ cells. As shown in Fig. 4a, overexpression of Rft1 resulted in a reduction of the relative levels of Man₃GlcNAc₂ with a concomitant increase in Man₇GlcNAc₂. This shift of the steady-state concentration of a cytoplasmically exposed oligosaccharide species to ER-lumenally exposed forms shows directly that overexpression of *RFT1* affects the topological distribution of the lipid-linked oligosaccharide.

We next examined asparagine-linked oligosaccharides in $\Delta alg11$ cells and $\Delta alg11$ cells that overexpress *RFT1*, after digesting protein fractions with peptide-N-glycosidase F (PNGaseF). Analysis of the released oligosaccharides revealed predominantly Man₆GlcNAc₂ in both strains (Fig. 4b), which probably results from the trimming of

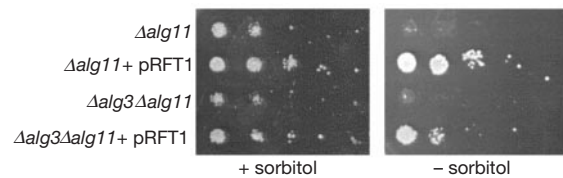


Figure 5 Overexpression of *RFT1* improves growth of $\Delta alg11$ (YG1365) and $\Delta alg3\Delta alg11$ (YG1363) strains. $\Delta alg11$ and $\Delta alg3\Delta alg11$ strains bearing either YEp352 or pJH01 (+pRFT1) were spotted onto SD-ura plates (with or without sorbitol) in rows. Each row consists of a serial dilution of indicated strain. The plates were incubated for 3 days at 30 °C.

protein-bound Man₇GlcNAc₂ by ER mannosidase^{14,21}. We conclude from these results that, as in wild-type cells, ER-lumenally oriented Man₇GlcNAc₂—but not cytoplasmically oriented Man₃GlcNAc₂—served as substrate for protein glycosylation. Moreover, overexpression of Rft1 in $\Delta alg11$ cells resulted in a shift towards CPY glycoforms containing more oligosaccharide units (Fig. 4b, compare lanes 3 and 5), demonstrating an improved glycosylation efficiency that is consistent with an increased pool of available substrate in the ER lumen. Notably, the oligosaccharides present on CPY molecules were sensitive to EndoH (Fig. 4b, lanes 3–6), supporting the conclusion that Man₇GlcNAc₂ and not Man₃GlcNAc₂ was transferred to protein. The increased glycosylation associated with *RFT1* overexpression also coincided with a partial suppression of the $\Delta alg11$ growth defect (Fig. 5). The overexpression of *RFT1* only altered the relative level of Man₇GlcNAc₂ to Man₃GlcNAc₂ oligosaccharide, but did not result in the appearance of a novel lipid-linked oligosaccharide.

To rule out the possibility that overexpression of *RFT1* has an effect on luminal mannose transferases, we constructed a $\Delta alg3\Delta alg11$ double mutant strain. In these cells, irrespective of *RFT1* overexpression, the lipid-linked Man₃GlcNAc₂ oligosaccharide was the largest oligonucleotide observed (data not shown).

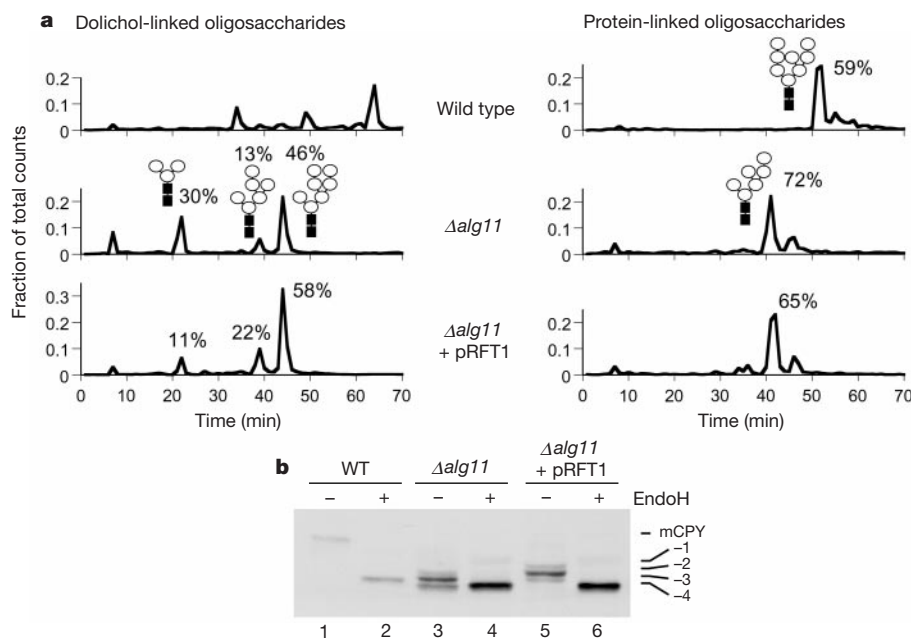


Figure 4 Overexpression of *RFT1* improves N-linked glycosylation in a $\Delta alg11$ strain. **a**, HPLC traces of [³H]-mannose-labelled lipid-linked oligosaccharides and N-linked oligosaccharides from wild-type (WT) cells and $\Delta alg11$ cells (YG1365) bearing either an empty plasmid ($\Delta alg11$) or *RFT1* on a multicopy plasmid ($\Delta alg11$ + pRFT1). Major HPLC peaks corresponding to Man₃GlcNAc₂, Man₆GlcNAc₂ and Man₇GlcNAc₂ of the lipid-linked oligosaccharide fraction, and Man₆GlcNAc₂ and Man₆GlcNAc₂ of N-linked oligosaccharide

fraction are marked. See Fig. 2b legend for definition of the symbols. In wild-type cells, lipid-linked Glc₃Man₃GlcNAc₂, Man₆GlcNAc₂ and Man₇GlcNAc₂ represent the predominant oligosaccharides. **b**, Proteins from the same cultures as **a** were extracted, incubated overnight in the presence (+) or absence (-) of EndoH, blotted, and the membranes probed with CPY antibody. Bands representing fully glycosylated (mCPY) and hypoglycosylated forms of CPY are indicated.

Again, we observed partial restoration of growth (see Fig. 5) and increased glycosylation of CPY (data not shown) on *RFT1* overexpression. These results further support the conclusion that Rft1 does not affect the structures of the lipid-linked oligosaccharides but rather their topological disposition in the ER membrane, thereby affecting the availability of the oligosaccharide substrate for N-glycosylation.

The experiments in this study support the hypothesis that Rft1 is required directly for membrane translocation of the Man₅GlcNAc₂ oligosaccharide across the ER membrane. The overexpression experiments in the *alg11* mutant (Fig. 4) suggest that Rft1 is the limiting component for the flipping reaction *in vivo* and that no additional factors may be required for its function. Rft1 is therefore an excellent candidate protein for the Man₅GlcNAc₂-PP-Dol flip-pase in the ER membrane. Of note, the results indicate that the translocation of Dol-P-Man^{7,22} across the ER membrane is independent of Rft1. Similar to the plasma membrane phospholipid scramblase²³, the fatty acid transporter FATP²⁴, and the bacterial Wzx protein²⁵, the amino-acid sequence of Rft1 does not reveal ATP-binding domains. Further genetic and biochemical experiments should allow us to address the mechanisms that make the translocation of polar and bulky substrates—such as the Man₅GlcNAc₂-PP-Dol intermediate—across the membrane possible, and whether membrane asymmetry is critical for the process. □

Methods

Yeast medium and methods

We used standard yeast medium and genetic techniques²⁶. With the exception of YG1137, all strains were grown on YPD medium unless otherwise stated. Strain YG1137 was maintained on YPGal. YG1363 and YG1365 were grown in medium supplemented with 1 M sorbitol unless otherwise stated.

Lipid- and protein-linked oligosaccharide analysis

Lipid-linked oligosaccharides were labelled, extracted and analysed as described²⁷. In brief, yeast cells (50 ml culture with an absorbance at 546 nm of 1) were grown in YPD and incubated in medium containing [³H]-mannose before lyses with organic solvents. Lipid-linked oligosaccharide was extracted using organic solvents and oligosaccharides were released by mild acid hydrolysis. The released oligosaccharides were analysed by HPLC using an NH₂ column with flow-through counting. We plotted the number of counts per minute divided by total counts in the run. The percentage of total signal in a sample is the average using two measurements. N-linked oligosaccharide was purified from cell debris after lipid-linked oligosaccharide extraction. Protein of the debris pellet was solubilized (10 min at 100 °C) in 0.2 ml 1% SDS, 50 mM Tris-HCl, 1% β-mercaptoethanol. After centrifugation (2 min at 15,000g) supernatant was supplemented to 1% (v/v) NP40 in 0.25 ml and protein-linked oligosaccharides were digested off using PNGaseF (2 units, overnight at 37 °C). Proteins were precipitated with 0.75 ml ethanol and samples were spun for 20 min at 15,000g. The supernatant was dried and resuspended in 0.2 ml 70:30 acetonitrile:water, 0.1 ml of which was analysed by HPLC as above.

Strain construction

The endogenous *RFT1* promoter was replaced with the *GAL1-10* promoter region using the polymerase chain reaction (PCR)-based method²⁸ in the background yeast strain SS328XS330, yielding strain YG1133 (*MATα ade2-201/ade2-201 ura3-52/ura3-52 his3Δ200/his3Δ200 tyr1/+ lys2-801/+ Pgal-rft1::KanMX/+*). YG1133 was sporulated and tetrads were dissected on YPGal to obtain a haploid strain with *RFT1* under the *GAL1-10* promoter control, YG1137 (*MATα ade2-201 ura3-52 his3Δ200 lys2-801 Pgal-rft1::KanMX*). Similarly, the entire *ALG11* open reading frame was replaced in SS328XS330 by integration of a PCR product containing the *S. cerevisiae HIS3* locus. Transformed yeast strain YG1141 (*MATα ade2-201/ade2-201 ura3-52/ura3-52 his3Δ200/his3Δ200 tyr1/+ lys2-801/+ Δalg11::HIS3/+*) was sporulated and tetrads were dissected to obtain a *Δalg11* haploid, YG1361 (*MATα ade2-201 ura3-52 his3Δ200 Δalg11::HIS3*), which was mated with YG248 (ref. 18). The resulting diploid YG1362 (*MATα ade2-201/ade2-201 ura3-52/ura3-52 his3Δ200/his3Δ200 lys2-801/+ Δalg3::HIS3 Δalg11::HIS3/+*) was sporulated on YPD plates containing 1 M sorbitol to obtain the haploid strains YG1365 (*MATα ade2-101 ura3-52 his3Δ200 Δalg11::HIS3*) and YG1363 (*MATα ade2-101 ura3-52 his3Δ200 lys2-801 Δalg3::HIS3 Δalg11::HIS3*). YG248 (*MATα Δalg3::HIS3 ade2-101 his3Δ200 lys2-801 ura3-52*)¹⁸ and the temperature-sensitive *dpm1-6* (*MATα dpm1::LEU2 leu2-3 lys2 ura3 trp1 pDM8-6*)¹⁹ strains were used.

Plasmid construction

To construct the Rft1 overexpression plasmid pJH01, the 2.8-kilobase (kb) *EcoRI*-*SphI* fragment containing the Rft1 locus of pDN387 (ref. 15) was ligated into the same sites of Yep352, a high-copy yeast plasmid.

Protein analysis

We performed protein extraction, EndoH treatment and western analysis as described¹⁸. Antibodies against CPY and chitinase were diluted 3,000- and 20,000-fold, respectively.

Received 30 August; accepted 16 November 2001.

- Kornfeld, R. & Kornfeld, S. Assembly of asparagine-linked oligosaccharides. *Annu. Rev. Biochem.* **54**, 631–664 (1985).
- Bugg, T. D. & Brandish, P. E. From peptidoglycan to glycoproteins: common features of lipid-linked oligosaccharide biosynthesis. *FEMS Microbiol. Lett.* **119**, 255–262 (1994).
- Higgins, C. F. Flip-flop: the transmembrane translocation of lipids. *Cell* **79**, 393–395 (1994).
- Sprong, H., van der Sluijs, P. & van Meer, G. How proteins move lipids and lipids move proteins. *Nature Rev. Mol. Cell Biol.* **2**, 504–513 (2001).
- Hetteima, E. H. & Tabak, H. F. Transport of fatty acids and metabolites across the peroxisomal membrane. *Biochim. Biophys. Acta* **1486**, 18–27 (2000).
- Borst, P., Zelcer, N. & van Helvoort, A. ABC transporters in lipid transport. *Biochim. Biophys. Acta* **1486**, 128–144 (2000).
- Abejón, C. & Hirschberg, C. B. Topography of glycosylation reactions in the endoplasmic reticulum. *Trends Biochem. Sci.* **17**, 32–36 (1992).
- Burda, P. & Aebi, M. The dolichol pathway of N-glycosylation. *Biochim. Biophys. Acta* **1426**, 239–257 (1999).
- Perez, M. & Hirschberg, C. B. Topography of glycosylation reactions in the rough endoplasmic reticulum membrane. *J. Biol. Chem.* **261**, 6822–6830 (1986).
- Snider, M. D. & Rogers, O. C. Transmembrane movement of oligosaccharide-lipids during glycoprotein synthesis. *Cell* **36**, 753–761 (1984).
- McCloskey, M. A. & Troy, F. A. Paramagnetic isoprenoid carrier lipids. 1. Chemical synthesis and incorporation into model membranes. *Biochemistry* **19**, 2056–2060 (1980).
- Hanover, J. A. & Lennarz, W. J. The topological orientation of N,N'-diacetylchitobiosylpyrophosphoryldolichol in artificial and natural membranes. *J. Biol. Chem.* **254**, 9237–9246 (1979).
- Huffaker, T. C. & Robbins, P. W. Yeast mutants deficient in protein glycosylation. *Proc. Natl. Acad. Sci. USA* **80**, 7466–7470 (1983).
- Cipollo, J. F., Trimble, R. B., Chi, J. H., Yan, Q. & Dean, N. The Yeast *ALG11* gene specifies addition of the terminal alpha 1,2-Man to the Man₅GlcNAc₂-PP-dolichol N-glycosylation intermediate formed on the cytosolic side of the endoplasmic reticulum. *J. Biol. Chem.* **276**, 21828–21840 (2001).
- Ng, D. T., Spear, E. D. & Walter, P. The unfolded protein response regulates multiple aspects of secretory and membrane protein biogenesis and endoplasmic reticulum quality control. *J. Cell Biol.* **150**, 77–88 (2000).
- Davidson, E. A. & Gowda, D. C. Glycobiology of *Plasmodium falciparum*. *Biochimie* **83**, 601–604 (2001).
- Hasilik, A. & Tanner, W. Carbohydrate moiety of carboxypeptidase Y and perturbation of its biosynthesis. *Eur. J. Biochem.* **91**, 567–575 (1978).
- Aebi, M., Gassenhuber, J., Domdey, H. & te Heesen, S. Cloning and characterization of the *ALG3* gene of *Saccharomyces cerevisiae*. *Glycobiology* **6**, 439–444 (1996).
- Orlean, P. Dolichol phosphate mannose synthase is required *in vivo* for glycosyl phosphatidylinositol membrane anchoring, O-mannosylation, and N-glycosylation of protein in *Saccharomyces cerevisiae*. *Mol. Cell Biol.* **10**, 5796–5805 (1990).
- Orlean, P., Kuranda, M. J. & Albright, C. F. Analysis of glycoproteins from *Saccharomyces cerevisiae*. *Methods Enzymol.* **194**, 682–697 (1991).
- Jelinek-Kelly, S. & Herscovics, A. Glycoprotein biosynthesis in *Saccharomyces cerevisiae*. Purification of the alpha-mannosidase which removes one specific mannose residue from Man₅GlcNAc. *J. Biol. Chem.* **263**, 14757–14763 (1988).
- Rush, J. S. & Wächter, C. J. Transmembrane movement of a water-soluble analogue of mannosyl-phosphoryldolichol is mediated by an endoplasmic reticulum protein. *J. Cell Biol.* **130**, 529–536 (1995).
- Zhou, Q. *et al.* Molecular cloning of human plasma membrane phospholipid scramblase. A protein mediating transbilayer movement of plasma membrane phospholipids. *J. Biol. Chem.* **272**, 18240–18244 (1997).
- Schaffer, J. E. & Lodish, H. F. Expression cloning and characterization of a novel adipocyte long chain fatty acid transport protein. *Cell* **79**, 427–436 (1994).
- Feldman, M. F. *et al.* The activity of a putative polyisoprenol-linked sugar translocase (Wzx) involved in *Escherichia coli* O antigen assembly is independent of the chemical structure of the O repeat. *J. Biol. Chem.* **274**, 35129–35138 (1999).
- Guthrie, C. & Fink, G. R. Guide to yeast genetics and molecular biology. *Methods Enzymol.* **194**, 3–37 (1991).
- Zufferey, R. *et al.* STT3, a highly conserved protein required for yeast oligosaccharyltransferase activity *in vitro*. *EMBO J.* **14**, 4949–4960 (1995).
- Wach, A. *et al.* in *Methods in Microbiology: Yeast Gene Analysis* (eds Tuite, M. & Brown, A. J.) 67–81 (Academic, San Diego, 1998).

Acknowledgements

We thank C. J. Waechter for his contributions. We also thank F. Reggiori and A. Conzelmann, and T. Immervoll and M. Gentzsch for providing antibodies against Gas1 protein and chitinase, respectively. P. Orlean provided the *dpm1-6* strain. This work was supported by grants from the Swiss National Science Foundation to M.A., the National Institutes of Health to P.W. and D.T.W.N., and the Canadian Institutes of Health Research to M.A.V. P.W. is an Investigator of the Howard Hughes Medical Institute.

Competing interests statement

The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to M.A. (e-mail: aebi@micro.biol.ethz.ch).

Characteristics of C₄ photosynthesis in stems and petioles of C₃ flowering plants

Julian M. Hibberd* & W. Paul Quick†

* Department of Plant Sciences, University of Cambridge, Downing Street, Cambridge CB2 3EA, UK

† Robert Hill Institute, Department of Animal and Plant Sciences, University of Sheffield, Western Bank, Sheffield S10 2TN, UK

Most plants are known as C₃ plants because the first product of photosynthetic CO₂ fixation is a three-carbon compound¹. C₄ plants, which use an alternative pathway in which the first product is a four-carbon compound, have evolved independently many times and are found in at least 18 families^{2,3}. In addition to differences in their biochemistry, photosynthetic organs of C₄ plants show alterations in their anatomy and ultrastructure⁴. Little is known about whether the biochemical or anatomical characteristics of C₄ photosynthesis evolved first. Here we report that tobacco, a typical C₃ plant, shows characteristics of C₄ photosynthesis in cells of stems and petioles that surround the xylem and phloem, and that these cells are supplied with carbon for photosynthesis from the vascular system and not from stomata. These photosynthetic cells possess high activities of enzymes characteristic of C₄ photosynthesis, which allow the decarboxylation of four-carbon organic acids from the xylem

and phloem, thus releasing CO₂ for photosynthesis. These biochemical characteristics of C₄ photosynthesis in cells around the vascular bundles of stems of C₃ plants might explain why C₄ photosynthesis has evolved independently many times.

Photosynthesis forms the basis on which all ecosystems on Earth function. The biochemical pathways of photosynthesis are highly conserved. Most plants are C₃ plants, in which the first product of photosynthesis is the three-carbon compound phosphoglyceric acid¹. A second biochemical pathway that allows the concentration of CO₂ in leaves has evolved in C₄ plants, which initially fix inorganic carbon in mesophyll cells into the four-carbon compound oxaloacetic acid. In C₄ plants, oxaloacetate is converted into malate or aspartate, which then diffuses into bundle sheath cells that surround the vascular bundles where decarboxylation supplies high concentrations of CO₂ to Rubisco. The C₄ pathway is thought to have evolved independently 31 times in 18 taxonomic families of angiosperms, and occurs in 8,000 to 10,000 species^{2,3}. The presence of C₄ photosynthesis in unicellular algae has also been reported⁵. In higher plants the C₄ pathway involves both biochemical and anatomical modifications, but it is not clear which of these modifications evolved first^{2,6}.

Some plants that possess characteristics of both C₃ and C₄ plants have been classified as C₃–C₄ intermediates⁷, and these may represent transitional stages in the evolution of C₄ photosynthesis from C₃ photosynthesis. In C₃–C₄ intermediates, a photorespiratory cycle has been reported to shunt the precursors of photorespiratory CO₂ from the mesophyll to the bundle sheath^{8,9}; in addition, it has been proposed that a pathway that recaptures CO₂ lost during photorespiration and that concentrates CO₂ around Rubisco in bundle sheath cells represents the first step in C₄ evolution^{6,7}.

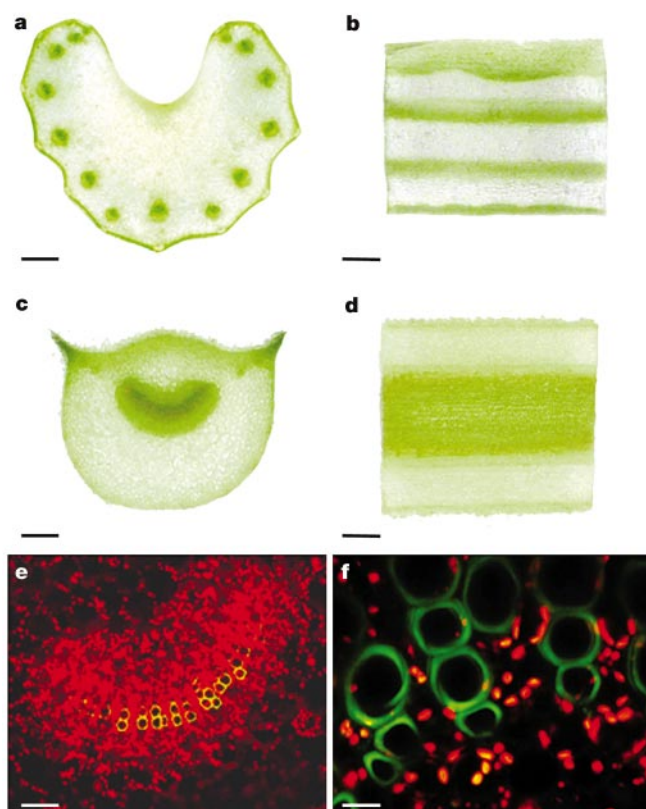


Figure 1 Distribution of chlorophyll in petioles of celery and tobacco. **a**, Transverse section of celery. **b**, Longitudinal section of celery. **c**, Transverse section of tobacco. **d**, Longitudinal section of tobacco. **e**, Confocal laser scanning microscope image showing chlorophyll fluorescence around the vascular system. **f**, Confocal laser scanning microscope image showing chloroplasts (red) in between the xylem vessels (green). Scale bars, 2.6 mm (**a**, **b**); 1.0 mm (**c**); 1.1 mm (**d**); 175 μ m (**e**); 25 μ m (**f**).

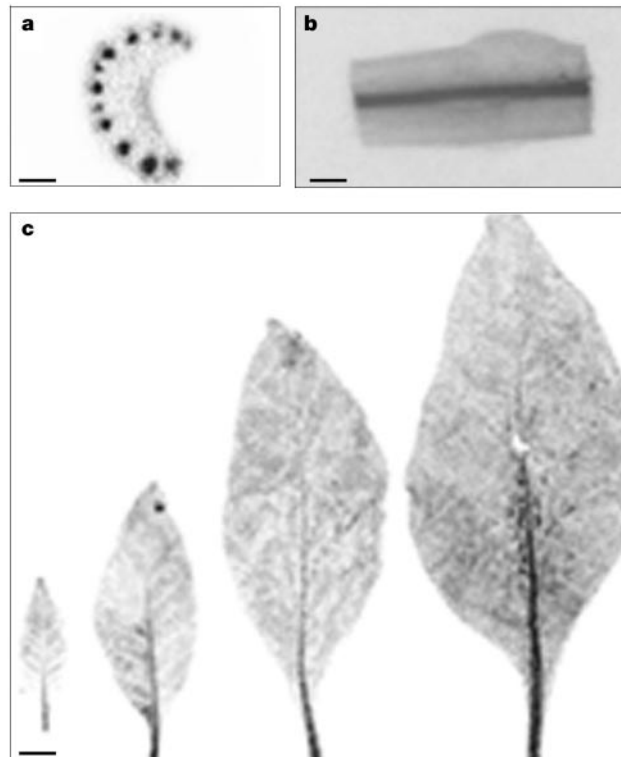


Figure 2 Incorporation of ¹⁴C into insoluble material in cells around the vascular system of celery and tobacco. **a**, Transverse section of celery after [¹⁴C]NaHCO₃ was fed to the xylem stream. Scale bar, 4.5 mm. **b**, Longitudinal section of tobacco showing ¹⁴C fixed into starch after [¹⁴C]NaHCO₃ was supplied to the xylem stream. Scale bar, 2.3 mm. **c**, Leaves from tobacco show that when malate with a ¹⁴C label on the C₄ position is supplied to the xylem stream, ¹⁴C is fixed in cells surrounding the vascular system. Young to old leaves are shown left to right. Scale bar, 14 mm.

Three separate biochemical pathways have evolved to decarboxylate the C_4 acid in the bundle sheath cells and thus supply Rubisco with CO_2 ; however, we know little about how these different pathways evolved. In addition to the limited number of biochemical pathways that allow fixation of inorganic carbon in flowering plants, there are few exceptions to the normal route that CO_2 takes to reach the sites of fixation. The pathway of CO_2 moving into a plant through stomata is almost universal in vascular land plants. Notable exceptions are submerged aquatic members of the Isoetaceae that do not produce stomata on submerged leaves and vary in their ability to form stomata on aerial leaves¹⁰. For example, the CAM (crassulacean acid metabolism) plant *Stylites andicola*, which is a rare ally of ferns, does not possess stomata and CO_2 is transferred to the leaves through its root system¹¹. In *Cuscuta reflexa*, a parasitic plant in which stomata have become modified into extrafloral nectaries, CO_2 fixed in photosynthesis is derived internally from respiration¹².

We studied a group of photosynthetic cells surrounding the vascular tissue of stems and petioles in tobacco that continue their association with vascular bundles well into the leaf tissue. Even in C_3 plants, cells around the veins are termed 'bundle sheath' cells^{13,14}. Although these photosynthetic cells are evident to the naked eye in plants such as celery (Fig. 1a, b), they have not been included in classifications of stem photosynthesis to date¹⁵. In transverse and longitudinal sections of celery (Fig. 1a, b) and tobacco (Fig. 1c–e), the most intense regions of chlorophyll accumulation are associated directly with internal vascular structures. Chlorophyll is found preferentially surrounding the xylem and the phloem of the vascular bundles (Fig. 1e), and chlorophyll fluorescence from the chloroplasts of cells directly between the xylem vessels can be observed (Fig. 1f). In fact, species from 30 different families that we have so far examined contain chlorophyll around the phloem and xylem vessels (data not shown).

In contrast to CO_2 movement to all other photosynthetic cells in flowering plants, CO_2 diffusion from stomata to the photosynthetic cells surrounding the vascular bundles is likely to be slow as there are relatively few stomata, many layers of surrounding cells and very few intercellular air spaces. This lack of air spaces between cells of the

vascular tissue has been previously noted and proposed to reduce diffusion of oxygen to cells that would be prone to oxidative stress¹⁶. But the photosynthetic cells that we observed have the potential to generate active oxygen species in cells close to the vascular system. As these photosynthetic cells are not close to stomata but are close to the xylem vessels, we tested whether they fix inorganic carbon supplied from the xylem stream.

Inorganic carbon supplied to roots and stems of tobacco and celery as [^{14}C]NaHCO₃ led to accumulation of ^{14}C in insoluble material, most notably starch, in cells associated with the vasculature (Fig. 2a, b). Supplying [^{14}C]glucose to roots (which is taken up by the roots and used in respiration to produce $^{14}CO_2$) led to ^{14}C in the xylem stream and the same distribution of isotope around the vascular bundles (data not shown). These data show that inorganic carbon surrounding roots can be taken up and fixed by cells around the xylem, and also that sugars used in respiration in roots give rise to CO_2 that moves into xylem vessels and is then fixed by cells surrounding the xylem. The concentration of CO_2 in the xylem sap is dependent on pH, and considerably higher than that in the air¹⁷. Supplying inorganic carbon to the roots of plants has been shown to enhance plant growth and productivity^{18,19}, although it should be noted that the growth enhancement is not completely due to fixation of the additional inorganic carbon supply^{19,20}.

The xylem stream of plants often contains organic acids^{21,22}, and the malate found in both phloem and xylem may have a role as a biochemical pH regulator to balance acid and base movement through the plant^{22,23}. We tested whether malate supplied to the transpiration stream would be decarboxylated, and the released CO_2 subsequently re-fixed photosynthetically by the cells surrounding the xylem vessels. Feeding malate labelled with ^{14}C in the C4 position (this is the carbon atom that is released during the decarboxylation process) to the xylem stream led to the fixation of ^{14}C in the veins of the petioles and stems of tobacco (Fig. 2c). Concentrations of malate in the xylem sap of tobacco plants decreased with increasing distances from the base of the plant (0.23 mM at the base to 0.02 mM in expanded leaves)²⁴. Changes in the concentration of malate in the xylem have implications for the acid–base balance of the shoot²², and therefore merit further

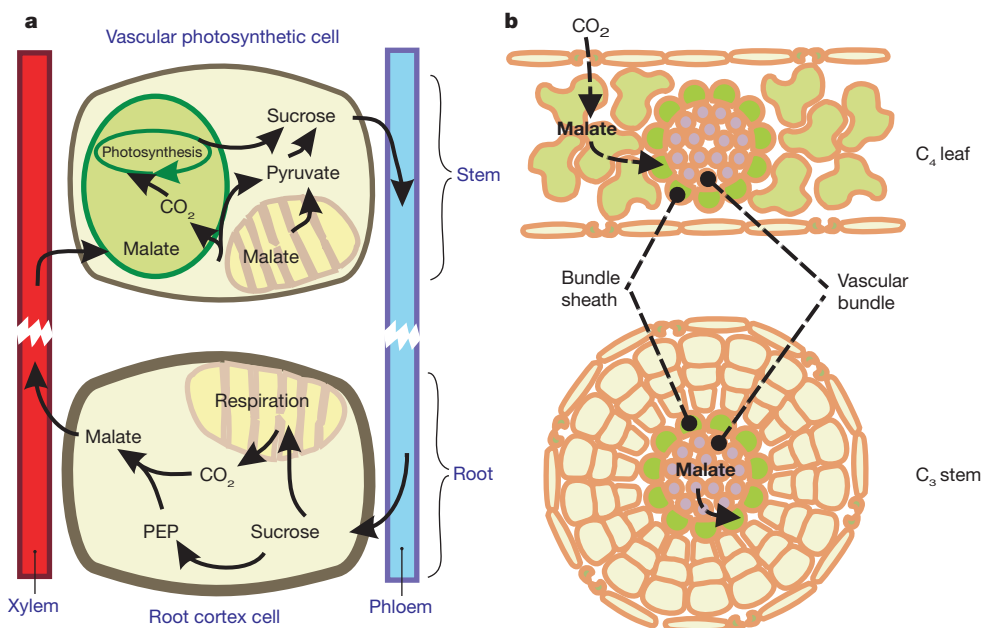


Figure 3 Photosynthesis around the veins of C_3 and C_4 plants. **a**, Model for the role of photosynthesis around the vascular system of stems and petioles of the C_3 plant tobacco. CO_2 is supplied from the soil, from respiration in roots or from malate that is produced in the root; either CO_2 or malate moves into the xylem stream, which supplies photosynthetic

cells lining the vascular system in stems and petioles. **b**, Comparison of the supply of malate to photosynthetic cells in a C_4 leaf and that to photosynthetic cells lining the vascular system in a C_3 stem. Both sets of cells that line the vascular system possess high activities of enzymes that can decarboxylate C_4 organic acids.

investigation. We conclude that these photosynthetic vascular cells scrub the xylem of malate, because the ^{14}C label was restricted to the petiole and basal regions of all leaves in a plant (Fig. 2c).

When the plants were left for 24 h (after supplying the transpiration stream with radiolabelled malate for 1 h), we found that ^{14}C no longer remained within the vascular tissues but was transferred to young developing leaves. Placing plants in darkness to close the stomata showed that this process was independent of the transpiration stream. We therefore conclude from this indirect data that sucrose produced from these cells around the vascular bundles is loaded into the phloem and translocated to growing apices. These data indicate that photosynthetic cells around the vasculature make use of organic acids present in the xylem of tobacco.

The decarboxylation of malate in cells around the veins of petioles and stems is reminiscent of the decarboxylation of organic acids in the bundle sheath of C_4 plants—that is, the cells that surround the vascular bundles of leaves. To test whether decarboxylation enzymes are specifically and preferentially expressed in cells around the vascular bundles in the stem of tobacco, we determined the activity of three decarboxylation enzymes—nicotinamide adenine dinucleotide (NAD)-malic, NADP-malic and phosphoenolpyruvate carboxykinase (PEPCK)—indicative of the three C_4 subtypes, and compared their activity in these cells with that in the leaf lamina. In cells surrounding the xylem and phloem of tobacco, the activity of NAD-malic enzyme (NAD-ME) was 13-fold higher than in the leaf, and the activity of NADP-malic enzyme (NADP-ME) and PEPCK were both 9-fold higher than in the leaf (Table 1). The activities of these decarboxylation enzymes far exceeds typical values obtained from C_3 plants and approaches activities expected in the bundle sheath of C_4 plants^{25,26}. These data indicate that in tobacco, a C_3 plant, similar biochemical attributes to those needed in C_4 photosynthesis are associated with photosynthesis around the vascular system of stems and petioles.

Because these enzymes are expressed highly and specifically in these cells, the correct regulatory elements controlling gene expression must be present. In agreement with our assays of enzyme activity, when present in tobacco the promoter of a gene coding NADP-ME from bean directs high and specific expression of the enzyme in cells surrounding the vascular tissue of stems²⁷. Notably, in the *Arabidopsis* genome the genes that code these decarboxylases are members of gene families that would allow the differential regulation and tissue-specific expression of certain members of each family. The vascular system may be regarded as taking the place of the mesophyll cells of a C_4 plant by supplying the surrounding photosynthetic cells with malate. The malate is then decarboxylated, and thus part of the C_4 cycle is operating in C_3 plants.

We also measured high activities of phosphoenol pyruvate orthophosphate dikinase (PPDK; this regenerates PEP from pyruvate in C_4 plants) in cells around the vascular system of tobacco petioles (Table 1). The role that the decarboxylases have in cells around the vascular system is not certain. It is possible that by controlling the concentration of malate, they are involved in controlling the pH of the xylem²². In addition, their high activity and the production of PEP through PPDK would allow for high provision of carbon skeletons into the shikimate pathway and thus

lignin synthesis in exactly the cells where lignin is incorporated into cell walls of the xylem vessels at high rates.

From these data we propose the following scheme. In C_3 plants, photosynthetic cells surrounding the vascular system are supplied with either malate or CO_2 from the xylem vessels (Fig. 3a). CO_2 can be supplied from root respiration, or enter the roots from the soil. Malate is produced from PEP because of the presence of phosphoenolpyruvate carboxylase in roots^{18,28,29}. Malate is then decarboxylated by cells bordering the vascular system in the stems of C_3 plants, and the CO_2 is used in photosynthesis to produce carbohydrates including sucrose and starch (Fig. 3a). The sucrose that is produced can be either used locally to fuel metabolism or loaded into the phloem and translocated to growing regions of the plant. When compared to the system that operates in C_4 plants, there are striking parallels (Fig. 3b). Bundle sheath cells of C_4 plants are photosynthetic, lie adjacent to vascular tissues, and possess high activities of enzymes that release CO_2 from malate. These key attributes of C_4 photosynthesis are also characteristics of the photosynthetic cells that surround the vascular bundle in C_3 plant stems and petioles that we have identified here. The main difference is the source of malate: in C_4 plants, malate is provided by adjacent mesophyll cells; in C_3 stems and petioles, malate is derived from the respiratory activity of distant heterotrophic tissues and transported through the vasculature to photosynthetic cells surrounding the vascular system.

We propose that the arrangement of photosynthetic cells around vascular bundles of C_3 stems is a more spatially separated version of the C_4 photosynthetic pathway. Furthermore, the presence of the correct biochemical pathways in cells adjacent to the vascular bundles in the stems and leaf veins of C_3 plants indicates that essential biochemical components and the regulatory elements controlling the cell-specific gene expression required for C_4 photosynthesis are already present in C_3 plants; thus, fewer modifications are needed for C_4 photosynthesis to evolve. The promoter of an NADP-ME gene from bean has been shown to direct expression in cells around the vascular system²⁷. The existence of such a similar pathway in C_3 plants provides a simple explanation for the polyphyletic evolution of C_4 photosynthesis. □

Methods

Plants and microscopy

Tobacco plants were grown under natural light in a greenhouse with supplementary lighting provided by metal halide bulbs supplying $250 \mu\text{mol photons m}^{-2} \text{s}^{-1}$. Plants were used during vegetative growth between 40 and 60 days after planting. We collected and analysed xylem sap as described²⁴. We cut sections for optical and confocal laser scanning microscopy manually with a razor blade. Images of chlorophyll fluorescence were taken using the TRITC channel of a Leica TCS confocal microscope.

^{14}C labelling and enzyme assays

We fed $1 \text{ mM } [^{14}\text{C}]\text{NaHCO}_3$ (specific activity $155 \text{ MBq mmol}^{-1}$), $0.1 \text{ mM } ^{14}\text{C}$ universally labelled glucose (74 MBq mmol^{-1}) or 0.2 mM malate (80 MBq mmol^{-1}) labelled in the C_4 position to the roots or stems of tobacco grown hydroponically, and to the stems of celery for 40 min in the light. Plants were left for 30 min, flash frozen, and heated in 70% ethanol to remove soluble sugars; sections were then exposed to X-ray film. We conducted enzyme assays on either leaf lamina of tobacco or the veins of tobacco petioles separated from the cortex manually with a razor blade; results are expressed per unit chlorophyll^{25,30}.

Received 13 September; accepted 21 November 2001.

Table 1 Activities of decarboxylation enzymes and PPDK

	NAD-ME activity ($\mu\text{mol min}^{-1}$ per mg Chl)	NADP-ME activity ($\mu\text{mol min}^{-1}$ per mg Chl)	PEPCK activity ($\mu\text{mol min}^{-1}$ per mg Chl)	PPDK activity ($\mu\text{mol min}^{-1}$ per mg Chl)
Leaf tissue	0.25 ± 0.03	0.78 ± 0.08	0.003 ± 0.0007	0.30 ± 0.05
Vein tissue	3.15 ± 0.22	6.82 ± 1.07	0.028 ± 0.009	5.33 ± 1.06
Relative enrichment (vein/leaf)	13-fold	9-fold	9-fold	18-fold

These enzymes are used by C_4 plants and are here measured in tobacco. Data shown are means $\pm$ s.e. ($n = 8-13$). Chl, chlorophyll.

- Calvin, M. *et al.* Carbon dioxide assimilation in plants. *Symp. Soc. Exp. Biol.* **5**, 284–305 (1951).
- Moore, P. D. Evolution of photosynthetic pathways in flowering plants. *Nature* **310**, 696 (1982).
- Sage, R. F., Li, M. & Monson, R. K. in *C₄ Plant Biology* (eds Sage, R. F. & Monson, R. K.) 551–584 (Academic, San Diego, 1999).
- Hatch, M. D. C_4 photosynthesis; a unique blend of modified biochemistry, anatomy and ultra-structure. *Biochem. Biophys. Acta* **895**, 81–106 (1987).
- Reinfelder, J. R., Kraepiel, A. M. L. & Morel, F. M. Unicellular C_4 photosynthesis in a marine diatom. *Nature* **407**, 996–999 (2000).
- Monson, R. K. in *C₄ Plant Biology* (eds Sage R. F. & Monson, R. K.) 377–410 (Academic, San Diego, 1999).
- Rawsthorne, S. C_3 – C_4 intermediate photosynthesis: linking physiology to gene expression. *Plant J.* **2**, 267–274 (1992).
- Rawsthorne, S., Hylton, C. M., Smith, A. M. & Woolhouse, H. W. Photorespiratory and immunogold localization of photorespiratory enzymes in leaves of C_3 and C_3 – C_4 intermediate species of *Morandia*. *Planta* **173**, 298–308 (1988).

9. Rawsthorne, S., Hylton, C. M., Smith, A. M. & Woolhouse, H. W. Distribution of photorespiratory enzymes between bundle sheath and mesophyll cells in leaves of the C_3 - C_4 intermediate species *Morinda arvensis* (L.) DC. *Planta* **176**, 527–532 (1988).
10. Keeley, J. E. CAM photosynthesis in submerged aquatic plants. *Bot. Rev.* **64**, 121–175 (1998).
11. Keeley, J. E., Osmond, C. B. & Raven, J. A. *Stylites*, a vascular land plant without stomata absorbs CO_2 via its roots. *Nature* **310**, 694–695 (1984).
12. Hibberd, J. M. et al. Localization of photosynthetic metabolism in the parasitic angiosperm *Cuscuta reflexa*. *Planta* **205**, 506–513 (1998).
13. Esau, K. *Plant Anatomy* 2nd edn (John Wiley, New York, 1965).
14. Kinsman, E. A. & Pyke, K. A. Bundle sheath cells and cell-specific plastid development in *Arabidopsis* leaves. *Development* **125**, 1815–1822 (1998).
15. Nilsen, E. T. in *Plant Stems* (ed. Gartner, B. L.) 223–240 (Academic, San Diego, 1995).
16. Raven, J. A. Long-term functioning of enucleate sieve elements: possible mechanisms of damage avoidance and damage repair. *Plant Cell Environ.* **14**, 139–146 (1991).
17. Gerendas, J. & Schurr, U. Physicochemical aspects of ion relations and pH regulation in plants—a quantitative approach. *J. Exp. Bot.* **50**, 1101–1114 (1999).
18. Cramer, M. D. & Richards, M. B. The effect of rhizosphere dissolved inorganic carbon on gas exchange characteristics and growth rates of tomato seedlings. *J. Exp. Bot.* **50**, 79–87 (1999).
19. Cramer, M. D., Gao, Z. F. & Lips, S. H. The influence of dissolved inorganic carbon in the rhizosphere on carbon and nitrogen metabolism in salinity treated tomato plants. *New Phytol.* **142**, 441–450 (1999).
20. Vuorinen, A. R., Rossi, P. & Vapaavuori, E. M. Combined effect of inorganic carbon and different nitrogen sources in the growth media on biomass production and nitrogen uptake in young willow and birch plants. *J. Plant Physiol.* **147**, 236–242 (1995).
21. Pate, J. S. in *Encyclopedia of Plant Physiology* Vol. 1 (eds Zimmerman, M. H. & Milburn, J. A.) 451–473 (Springer, Heidelberg, 1975).
22. Raven, J. A. & Smith, F. A. Nitrogen assimilation and transport in vascular land plants in relation to intracellular pH regulation. *New Phytol.* **76**, 415–431 (1976).
23. Ben Zioni, A., Vaadia, Y. & Lips, S. H. Nitrate uptake by roots as regulated by nitrate reduction products of the shoot. *Physiol. Plant.* **24**, 288–290 (1971).
24. Hibberd, J. M., Quick, W. P., Press, M. C., Scholes, J. D. & Jeschke, W. D. Solute fluxes from tobacco to the parasitic angiosperm *Orobancha cernua* and the influence of infection on host carbon and nitrogen relations. *Plant Cell Environ.* **22**, 937–947 (1999).
25. Ashton, A. R., Burnell, J. N., Furbank, R. T., Jenkins, C. L. D. & Hatch, M. D. in *Methods in Plant Biochemistry*, Vol. 3 (ed. Lea, P. J.) 39–72 (Academic, San Diego, 1990).
26. Lea, P. J. & Leegood, R. C. *Plant Biochemistry and Molecular Biology* (Wiley, Chichester, 1999).
27. Schaaf, J., Walter, M. H. & Hess, D. Primary metabolism in plant defence. *Plant Physiol.* **108**, 949–960 (1995).
28. Gao, Z., Sagi, M. & Lips, M. Assimilate allocation priority as affected by nitrogen compounds in the xylem sap of tomato. *Plant Physiol. Biochem.* **34**, 807–815 (1996).
29. Thomas, J. C., De Armond, R. L. & Bohnert, H. J. Influence of NaCl on growth, proline and phosphoenolpyruvate carboxylase levels in *Mesembryanthemum crystallinum* suspension cultures. *Plant Physiol.* **98**, 626–631 (1992).
30. Walker, R. P., Trevanion, S. J. & Leegood, R. C. Phosphoenolpyruvate carboxykinase from higher plants: Purification from cucumber and evidence of rapid proteolytic cleavage in extracts from a range of plant tissues. *Planta* **196**, 58–63 (1995).

Acknowledgements

We thank J. C. Gray and W. D. Jeschke for discussions. J.M.H. thanks the BBSRC for a Sir David Phillips Research Fellowship.

Competing interests statement

The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to J.M.H. (e-mail: julian.hibberd@plantsci.cam.ac.uk).

Anaerobic microbial metabolism can proceed close to thermodynamic limits

Bradley E. Jackson & Michael J. McInerney

Department of Botany and Microbiology, University of Oklahoma,
770 Van Vleet Oval, Norman, Oklahoma 73019-2045, USA

Many fermentative bacteria obtain energy for growth by reactions in which the change in free energy ($\Delta G'$) is less than that needed to synthesize ATP^{1-4} . These bacteria couple substrate metabolism directly to ATP synthesis, however, by classical phosphoryl transfer reactions^{4,5}. An explanation for the energy economy of these organisms is that biological systems conserve energy in discrete

amounts^{3,4}, with a minimum, biochemically convertible energy value of about -20 kJ mol^{-1} (refs 1–3). This concept predicts that anaerobic substrate decay ceases before the minimum free energy value is reached, and several studies support this prediction^{1,6–9}. Here we show that metabolism by syntrophic associations, in which the degradation of a substrate by one species is thermodynamically possible only through removal of the end product by another species¹, can occur at values close to thermodynamic equilibrium ($\Delta G' \approx 0 \text{ kJ mol}^{-1}$). The free energy remaining when substrate metabolism halts is not constant; it depends on the terminal electron-accepting reaction and the amount of energy required for substrate activation. Syntrophic associations metabolize near thermodynamic equilibrium, indicating that bacteria operate extremely efficient catabolic systems.

Anaerobic bacteria capable of syntrophic metabolism thrive in environmental niches where amounts of free energy are at a minimum¹, and therefore represent ideal microbial systems for testing bioenergetic concepts. In syntrophic associations, the degradation of a substrate by one species is made thermodynamically possible through the removal of end products by another species (ref. 1; and Table 1). Syntrophic interactions are widespread, occurring in both manmade and natural environments such as sewage digesters, municipal landfills, hydrocarbon-contaminated soils, freshwater sediments and waterlogged soils^{1,10–12}. The ubiquity of such interactions emphasizes the fact that, in most environments, metabolic cooperation among microbial species is required for the complete destruction of organic matter^{1,12}.

We observed previously that syntrophic cultures only partially degrade their substrate, stopping at a threshold concentration^{7,13}. The resulting substrate threshold was large in some cases, up to 1.5 mM, and remained unchanged for up to 360 days of incubation¹³. Experiments have shown that nutritional limitation, loss of metabolic activity, inhibition by high hydrogen or formate concentrations, kinetic inhibition and toxicity from build-up of the undissociated form of acetate are not the cause of this threshold^{7,8,13}. The substrate threshold concentration was, however, influenced by the accumulation of end product (specifically acetate). We concluded that the extent of substrate degradation was controlled by a thermodynamic, mass-action effect^{7,13}.

By manipulating substrate degradation and threshold formation, we tested the hypothesis that a minimum quantum of free energy is needed to sustain microbial metabolism. On the basis of universal biological energy quantum theory^{1–3}, we posited the following predictions: first, substrate degradation will cease at a threshold concentration that corresponds to a critical free energy value ($\Delta G'_{\text{crit}}$) independent of the species or substrate involved; second, the $\Delta G'_{\text{crit}}$ value at threshold will be constant (not appreciably lower or higher than -20 kJ mol^{-1}); and third, the $\Delta G'_{\text{crit}}$ will not be affected by different substrates or terminal electron-accepting schemes.

Benzoate and butyrate were used as model substrates, because each of these is a key intermediate in the degradation of diverse organic compounds^{3,12,14}. Each compound was added to defined cocultures containing a syntrophically oxidizing bacterium and a hydrogen-consuming partner. Table 2 shows the results of butyrate degradation by a typical *Syntrophus aciditrophicus* coculture grown under sulphate-reducing conditions (with *Desulfovibrio* sp. strain G11) when stressed with different concentrations of acetate. Roughly 2.5 mM butyrate was degraded incompletely to a threshold concentration that ranged from 5 to 250 μM , depending on the final concentration of acetate. Hydrogen partial pressures were low, ranging from 0.22 ± 0.06 to $0.31 \pm 0.05 \text{ Pa}$, and formate was not detected (detection limit of 1 μM).

The extent of butyrate degradation depended on the amount of acetate in the system, which is consistent with the interpretation that the accumulation of end products thermodynamically controls the extent of substrate degradation. Even though the final butyrate

Table 1 Reactions involved in syntrophic metabolism

Reactions	$\Delta G^{\circ\prime}$ (kJ per reaction)	$\Delta G^{\prime\prime}$ (kJ per reaction)
Methanogenic H_2 consumption		
$4 H_2 + HCO_3^- + H^+ \rightarrow CH_4 + 3 H_2O$	-135.6	-32.7
Syntrophic degradation without H_2 consumption		
$Butyrate^- + 2 H_2O \rightarrow 2 acetate^- + H^+ + 2 H_2$	+48.3	NA†
$Benzoate^- + 7 H_2O \rightarrow 3 acetate^- + HCO_3^- + 3 H^+ + 3 H_2$	+70.6	NA
Syntrophic degradation with H_2 consumption		
$2 butyrate^- + HCO_3^- + H_2O \rightarrow 4 acetate^- + H^+ + CH_4$	-39.4	-50.0
$4 benzoate^- + 19 H_2O \rightarrow 12 acetate^- + HCO_3^- + 9 H^+ + 3 CH_4$	-124.4	-289.9

Respective Gibbs free energy changes are given for the reactions at standard and physiological conditions.

* Calculated from data in ref. 5 with the free energy of formation for benzoate given in ref. 30.

† Calculated on the basis of the following concentrations observed for methanogenic conditions: p_{H_2} , 10 Pa; p_{CH_4} , 0.7×10^5 Pa; butyrate, 10 μ M; benzoate, 3 μ M; acetate, 100 μ M; bicarbonate, 35 mM.

‡ NA, not applicable.

threshold concentrations differed between acetate treatments by nearly two orders of magnitude, the $\Delta G'$ value was remarkably constant for all conditions (Table 2). The average $\Delta G'$ for butyrate degradation was -13.8 ± 1.2 kJ mol⁻¹; however, this value differed from that expected according to biological energy quantum theory. When a different primary partner was used, *Syntrophomonas wolfei* in coculture with the same sulphate-reducing bacterium (*Desulfovibrio* strain G11), the average $\Delta G'_{crit}$ for butyrate degradation was -12.7 ± 2.0 kJ mol⁻¹—a value also lower than predicted. This showed that $\Delta G'_{crit}$ was independent of the primary fermenting species. These results and previous reports of low or variable Gibbs free energies^{6,15–17} prompted us to test whether the universal biological energy quantum theory accurately describes experimental data obtained with a variety of microorganisms, substrates and electron-accepting schemes.

For syntrophic butyrate degradation, the average $\Delta G'$ was -4.5 ± 1.9 under methanogenic conditions (*S. wolfei* or *S. aciditrophicus* in coculture with *Methanospirillum hungatei*), -13.1 ± 1.6 under sulphate-reducing conditions (*S. wolfei/Desulfovibrio* strain G11 coculture) and -17.5 ± 0.5 under nitrate-reducing conditions (*S. wolfei/Desulfovibrio* strain G11 coculture). For syntrophic benzoate degradation, the average $\Delta G'$ was -24.7 ± 3.3 under methanogenic conditions (*S. aciditrophicus* or *Syntrophus gentianae* in coculture with *M. hungatei*), -33.4 ± 3.5 under sulphate-reducing conditions (*S. aciditrophicus* or *Syntrophus buswellii* in coculture with *Desulfovibrio* strain G11) and -45.1 ± 4.3 under nitrate-reducing conditions (*S. aciditrophicus/Desulfovibrio* strain G11 coculture).

Analysis of variance showed that the means for the $\Delta G'$ at threshold with different acetate concentrations or different species for a given substrate and electron-accepting condition were not statistically different ($P < 0.001$), whereas the means for the $\Delta G'$ at threshold for each electron-accepting condition for each substrate were statistically different from each other. Thus, the $\Delta G'$ value for threshold formation for syntrophic butyrate or benzoate degradation was the same for a given terminal electron-accepting condition (sulphate-reducing, nitrate-reducing or methanogenic), irrespective of the primary catabolic partner (that is, *S. wolfei*, *S. aciditrophicus*, *S. gentianae* or *S. buswellii*). The $\Delta G'$ for syntrophic butyrate degradation under methanogenic conditions was very low but

consistent with a thermodynamically controlled threshold value. The hydrogen partial pressure ranged from 1.2 ± 0.1 to 1.9 ± 0.3 Pa. All final butyrate concentrations were within 5 μ M of their respective values under sulphate-reducing conditions (Table 2), apart from a final butyrate concentration of 106.0 ± 14.2 μ M, which was observed under methanogenic conditions when 60 mM acetate was added.

The difference in $\Delta G'$ values observed at threshold between syntrophic butyrate and benzoate metabolism is probably related to the amount of energy required for butyrate and benzoate activation, respectively. All members of the genus *Syntrophus* use the enzyme benzoyl-CoA ligase, which requires the stoichiometric conversion of ATP to AMP and pyrophosphate for the activation of benzoate^{14,18–20}. This is in contrast to the activation strategy used by *S. wolfei*, which uses an energy-neutral, butyryl-acetyl-CoA transferase to activate butyrate²¹. Consistent with this hypothesis was the finding that the $\Delta G'$ value obtained at threshold for syntrophic butyrate metabolism by sulphate-reducing cocultures of *S. aciditrophicus* and *Desulfovibrio* strain G11 (-13.8 ± 1.2 kJ mol⁻¹) was not statistically different from that observed for cocultures of *S. wolfei* and *Desulfovibrio* strain G11 (-12.7 ± 2.0 kJ mol⁻¹). When *S. aciditrophicus* was grown with benzoate with the same sulphate reducer and sulphate, however, the $\Delta G'$ value at threshold was much higher (near -30 kJ mol⁻¹).

From a theoretical standpoint, two points can be raised regarding the assumptions used to formulate the current 'minimum quantum of energy' theory. First, bacteria require about 70 kJ mol⁻¹ of free energy to conserve 1 mol of ATP under the physiological conditions that are presumed to prevail in a cell. This assumption is made on the basis of actively growing *Escherichia coli* cells (ATP, 10 mM; ADP, 1 mM; inorganic phosphate, 10 mM), which may not represent the phosphorylation potential of a bacterium nearing its thermodynamic limits for substrate metabolism. It has been shown that the ADP/ATP ratio changes with the metabolic state of the cell, possibly lowering the amount of free energy required to generate ATP^{22,23}.

Second, the stoichiometry for ATP synthesis through ion translocation is a fixed ratio²⁴. If the stoichiometry of ion transported to ATP formed is different from a ratio of 3 to 1, or if the ratio varies according to the energetic state of the cell—for example, growing versus non-growing conditions—then a constant energy quantum may not be required as the energetic limits of substrate metabolism are reached. Variable stoichiometries of 2–4 mol H⁺ per mol ATP have been reported for ATP-driven proton pumps in *E. coli*²⁵. On the basis of crystalline structures of the F_0 subunit of a *Saccharomyces cerevisiae* proton-driven ATP synthase, the H⁺/ATP ratio is non-integral and its value lies between 3 and 4 (ref. 26).

Our experimental observations imply that the above assumptions may not be universally appropriate, especially for anaerobic bacteria, and must be reassessed to describe correctly the energetic processes of diverse bacterial metabolisms. To our knowledge, this is the first body of research that has tested the validity of the biological

Table 2 Effect of acetate on butyrate degradation by *S. aciditrophicus*

Addition*	Final acetate (mM)	Final butyrate (μ M)	$\Delta G'$ (kJ mol ⁻¹)
None	5.1 ± 0.4	4.7 ± 1.6	-14.7 ± 0.5
10 mM acetate	14.0 ± 0.3	13.9 ± 1.0	-13.8 ± 0.2
30 mM acetate	34.6 ± 1.4	50.9 ± 6.8	-12.5 ± 0.5
60 mM acetate	64.0 ± 1.6	251 ± 87	-12.7 ± 1.2
60 mM NaCl	5.2 ± 0.6	5.0 ± 1.9	-15.2 ± 0.7

Cocultures of *S. aciditrophicus* were grown under sulphate-reducing conditions. The effects of acetate on the thermodynamics and threshold value of butyrate degradation are shown. Values are the means $\pm$ s.d. of triplicate determinations.

* All tubes contained roughly 2.5 mM butyrate.

energy quantum hypothesis by systematically investigating varied species, substrates and electron-accepting schemes at the point where metabolism ceases and a thermodynamic equilibrium is established. Syntrophically metabolizing bacteria may maximize energy conservation when thermodynamic constraints begin to limit substrate degradation, possibly by altering their phosphorylation potential or electromotive membrane potential in response to signals from their hydrogen-using partner. Our results indicate that the free energy remaining when metabolism stops is highly variable, and imply that the extent of metabolism depends on the activation steps used for substrate metabolism as well as the terminal electron-accepting condition. Collectively, these data show that a universal-minimum free energy does not predict the extent of substrate metabolism and that bacterial metabolism can proceed near thermodynamic equilibrium—a condition that is often thought to be a biological improbability. □

Methods

Cultivation

The methods for the preparation and use of anaerobic media and solutions have been described²⁷. We used a basal medium²⁸ with 7 mM sodium benzoate or 10 mM sodium butyrate and 0.05% clarified rumen fluid for all cocultures. Sodium sulphate and sodium nitrate were added to a final concentration of 10 mM. All cultures were grown at 35 °C except for *S. gentianae*, which was grown at 30 °C. We grew *S. wolfei* in coculture with either *M. hungatei* in the butyrate basal medium, or *Desulfovibrio* sp. strain G11 in the butyrate basal medium with sulphate or nitrate. *S. aciditrophicus* was grown in coculture either with *M. hungatei* in the benzoate basal medium or with *Desulfovibrio* sp. strain G11 in the benzoate basal medium with sulphate or nitrate added. We grew *S. aciditrophicus* in butyrate basal medium with sulphate in coculture with *Desulfovibrio* sp. strain G11. *S. gentianae* was grown in benzoate basal medium in coculture with *M. hungatei*, and *S. buswellii* was grown in benzoate basal medium with sulphate in coculture with *Desulfovibrio* strain G11.

Substrate degradation

We monitored metabolism until a stable substrate threshold was observed, typically 45–90 days. From the direct measurement of all substrates and products, the energetic status of each culture was assessed by determining the change in free energy (ΔG°) for substrate degradation at the point where a stable threshold concentration was established^{1,7,13}. We used the Nernst equation to correct Gibbs energy values for the actual temperature, and the Van't Hoff equation to correct the free energy of formation of reactants and products⁴. Correction for ionic strength was calculated using the Debye–Hückel equation and had a negligible effect on Gibbs energy values (0.15 kJ mol⁻¹). Values are reported as the mean $\pm$ s.d. ($n = 3–10$). Analysis of variance (ANOVA) values were determined using the statistical functions in Microsoft Excel.

We measured benzoate using high-performance liquid chromatography, with a reverse phase C₁₈ column and detection at 254 nm. A flow rate of 1.0 ml min⁻¹ was used with a mobile phase of 80% (v/v) 50 mM sodium acetate (pH 4.5) and 20% (v/v) acetonitrile. The benzoate detection limit was roughly 250 nM. Formate was measured using a Dionex DX-500 ion chromatography system, with an AS-11 anion exchange column and a CD 20 conductivity detector. A flow rate of 1.5 ml min⁻¹ was used with a mobile phase of 0.5 mM NaOH. We also measured formate by monitoring at 320 nm formation of the reduced form of nicotinamide adenine dinucleotide (NADH) from the enzymatic reduction of NAD by formate dehydrogenase⁷. The detection limit for both methods was 1 μ M. Butyrate and acetate were measured by gas chromatography, with a flame ionization detector (FID) and a 1.8-m glass column packed with Carbowax B DA 80/20 4% Carbowax 20M resin. The column temperature was increased from 165 to 185 °C over 3.5 min, with an injector temperature of 180 °C and a detector temperature of 200 °C. We used a flow rate of 24 ml min⁻¹ helium.

Carbon dioxide was quantified by gas chromatography, with a thermal conductivity detector and a 1.8-m stainless steel column packed with Porapak Q resin. A flow rate of 20 ml min⁻¹ helium was used, with an injector temperature of 100 °C, an oven temperature of 50 °C and at a filament temperature of 180 °C. Methane was quantified by FID gas chromatography, with a 1.8-m stainless steel column packed with Porapak Q resin. A flow rate of 30 ml min⁻¹ helium was used with an injector temperature of 100 °C, an oven temperature of 60 °C and a detector temperature of 125 °C. We quantified hydrogen with a mercury vapour reduction gas analyser²⁹.

Received 6 September; accepted 21 November 2001.

1. Schink, B. Energetics of syntrophic cooperation in methanogenic degradation. *Microbiol. Mol. Biol. Rev.* **61**, 262–280 (1997).

2. Schink, B. in *Biotechnology, Focus 2* (eds Finn, R. K. & Präve, P.) 63–89 (Hanser, Munich, 1990).
3. Schink, B. & Thauer, R. in *Granular Anaerobic Sludge: Microbiology and Technology* (eds Lettinga, G. et al.) 5–17 (Pudoc, Wageningen, 1988).
4. Thauer, R. K. & Morris, J. G. Metabolism of chemotrophic anaerobes: old views and new aspects. *Symf. Soc. Gen. Microbiol.* **36**, 123–168 (1984).
5. Thauer, R. K., Jungermann, K. & Decker, K. Energy conservation in chemotrophic anaerobic bacteria. *Bacteriol. Rev.* **41**, 100–180 (1977).
6. Kleerebezem, R. & Stams, A. J. M. Kinetics of syntrophic cultures: A theoretical treatise on butyrate fermentation. *Biotechnol. Bioeng.* **67**, 529–543 (2000).
7. Warikoo, V., McNerney, M. J., Robinson, J. A. & Sullita, J. M. Interspecies acetate transfer influences the extent of anaerobic benzoate degradation by syntrophic consortia. *Appl. Environ. Microbiol.* **62**, 26–32 (1996).
8. Schöcke, L. & Schink, B. Energetics of methanogenic benzoate degradation by *Syntrophus gentianae* in syntrophic coculture. *Microbiology* **143**, 2345–2351 (1997).
9. Christensen, T. H. et al. Characterization of redox conditions in groundwater contaminant plumes. *J. Contam. Hydrol.* **45**, 165–241 (2000).
10. Coutts, D. A. P., Senior, E. & Balba, M. T. M. Multi-stage chemostat investigation of interspecies interactions in a hexanoate-catabolizing microbial association isolated from anoxic landfill. *J. Appl. Bacteriol.* **62**, 731–740 (1987).
11. Dojka, M. A., McNerney, M. J., Haack, S. K. & Pace, N. R. Microbial diversity in a hydrocarbon- and chlorinated-solvent-contaminated aquifer undergoing intrinsic bioremediation. *Appl. Environ. Microbiol.* **64**, 3869–3877 (1998).
12. McNerney, M. J. & Bryant, M. P. in *Biomass Conversion Processes for Energy and Fuels* (eds Sofer, S. S. & Zaborsky, O. R.) 277–296 (Plenum, New York, 1981).
13. Jackson, B. E. Thermodynamics and energetics of syntrophic substrate degradation. Thesis, Univ. Oklahoma (1999).
14. Harwood, C. S., Burchardt, G., Hermann, H. & Fuchs, G. Anaerobic metabolism of aromatic compounds via the benzoyl-CoA pathway. *FEMS Microbiol. Rev.* **22**, 439–458 (1999).
15. Seitz, H.-J., Schink, B., Pfennig, N. & Conrad, R. Energetics of syntrophic ethanol oxidation in defined chemostat cocultures. I. Energy requirement for H₂ production and H₂ oxidation. *Arch. Microbiol.* **155**, 82–88 (1990).
16. Dwyer, D. F., Weeg-Aerenssens, E., Shelton, D. R. & Teidje, J. M. Bioenergetic conditions of butyrate metabolism by a syntrophic, anaerobic bacterium in coculture with hydrogen-oxidizing methanogenic and sulfidogenic bacteria. *Appl. Environ. Microbiol.* **54**, 1354–1359 (1988).
17. Scholten, J. C. M. & Conrad, R. Energetics of syntrophic propionate oxidation in defined batch and chemostat cocultures. *Appl. Environ. Microbiol.* **66**, 2934–2942 (2000).
18. Auburger, G. & Winter, J. Purification and characterization of the benzoyl-CoA ligase from a syntrophic, benzoate-degrading anaerobic mixed culture. *Appl. Microbiol. Biotechnol.* **37**, 789–795 (1992).
19. Elshahed, M. S., Bhupathiraju, V. K., Wofford, N. Q., Nanny, M. A. & McNerney, M. J. Metabolism of benzoate, cyclohex-1-ene carboxylate and cyclohexane carboxylate by *Syntrophus aciditrophicus* strain SB in syntrophic association with H₂-using microorganisms. *Appl. Environ. Microbiol.* **67**, 1728–1738 (2001).
20. Harwood, C. S. & Gibson, J. Shedding light on anaerobic benzene ring degradation: a process unique to prokaryotes? *J. Bacteriol.* **179**, 301–309 (1997).
21. Wofford, N. Q., Beaty, P. S. & McNerney, M. J. Preparation of cell-free extracts and the enzymes involved in fatty acid metabolism in *Syntrophomonas wolfei*. *J. Bacteriol.* **167**, 179–185 (1986).
22. Tran, Q. H. & Uden, G. Changes in the proton potential and the cellular energetics of *Escherichia coli* during growth by aerobic and anaerobic respiration or by fermentation. *Eur. J. Biochem.* **251**, 538–543 (1998).
23. Jetten, M. S. M., Stams, A. J. M. & Zehnder, A. J. B. Adenine nucleotide content and energy charge of *Methanohalobium salinarum* during acetate degradation. *FEMS Microbiol. Lett.* **84**, 313–318 (1991).
24. Maloney, P. Relationship between phosphorylation potential and electrochemical H⁺ gradient during glycolysis in *Streptococcus lactis*. *J. Bacteriol.* **153**, 1461–1470 (1983).
25. Cramer, W. A. & Knaff, D. B. *Energy Transduction in Biological Membranes: a Textbook of Bioenergetics* (Springer, New York, 1990).
26. Stock, D., Leslie, A. G. W. & Walker, J. E. Molecular architecture of the rotary motor in ATP synthase. *Science* **286**, 1700–1705 (1999).
27. Balch, W. E. & Wolfe, R. S. New approach to the cultivation of methanogenic bacteria: 2-mercaptoethanesulfonic acid (HS-CoM)-dependent growth of *Methanobacterium ruminantium* in a pressurized atmosphere. *Appl. Environ. Microbiol.* **32**, 781–791 (1976).
28. McNerney, M. J., Bryant, M. P. & Pfennig, N. An anaerobic bacterium that degrades fatty acids in syntrophic association with methanogens. *Arch. Microbiol.* **122**, 129–135 (1979).
29. Seiler, W., Giehl, H. & Roggendorf, P. Detection of carbon monoxide and hydrogen by conversion of mercury oxide to mercury vapor. *Atmos. Technol.* **12**, 40–45 (1980).
30. Kaiser, J. L. & Hanselmann, K. W. Fermentative metabolism of substituted monoaromatic compounds by a bacterial community from anaerobic sediments. *Arch. Microbiol.* **133**, 384–391 (1982).

Acknowledgements

This work was supported by a grant from the US Department of Energy.

Competing interests statement

The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to M.J.M. (e-mail: mcinerney@ou.edu).

Contacts

Publisher: Fabien Savenay
Editor: Paul Smaglik
Sales Director: Ben Crowe

European Head Office, London

The Macmillan Building
4 Crinan Street
London N1 9XW, UK
Tel +44 (0) 20 7843 4961
Fax +44 (0) 20 7843 4996
e-mail: naturejobs@nature.com

Group European Manager:
Leonie Welss (4954)
European Manager:
Nevin Bayoumi (4978)

UK/ RoW/ Ireland:

Matt Powell (4953), Ben Corp (4974),
Andy Douglas (4975)

Holland/ Italy: Nevin Bayoumi (4978)

Scandinavia: Silke Opstrup (4994)

Spain/ Portugal: Leonie Welss
(4954)

Production Manager:

Billie Franklin
To send films and materials use
London address above.
Tel +44 (0) 20 7843 4814
Fax +44 (0) 20 7843 4996
e-mail: naturejobs@nature.com

International

Advertising Coordinator:

Laura Pearson (4977)

Naturejobs web development:

Tom Hancock

Naturejobs online production:

Ben Lund

European Satellite Offices

France/ Belgium:

Christine Niox-Chateau
Tel + 33 (0) 1 43 20 16 51
Fax + 33 (0) 1 43 20 51 52
e-mail: c.nioxchateau@nature.com

Germany/ Austria/ Switzerland:

Patrick Phelan/ Kate Turner
Tel + 49 89 54 90 57 11/-2
Fax + 49 89 54 90 57 20
e-mails: p.phelan@nature.com
k.turner@nature.com

US Head Office, New York

345 Park Avenue South,
10th Floor, New York, NY 10010-1707
Tel +1 800 989 7718
Fax +1 800 989 7103
e-mail: naturejobs@nature.com

US Sales Director:

Ben Crowe

US Sales Manager:

Peyton Mason

US Advertising Coordinator:

Ashly de Leon

Japan Head Office, Tokyo

MG Ichigaya Building (5F),
19-1 Harakatomachi,
Shinjuku-ku,
Tokyo 162-0841
Tel +81 3 3267 8751
Fax +81 3 3267 8746
e-mail: k.cowan@naturejp.com
Japan Manager: Kate Cowan

naturejobs

Avoiding 'extinct volcanoes'

Six months after becoming president at Harvard University, Lawrence Summers is making ripples in faculty recruitment and retention that are attracting attention beyond the confines of the campus. He is searching for faculty members who he thinks have their best work ahead of them and is avoiding more experienced academics whose future may not be so productive — people who Jeremy Knowles, Harvard dean of arts and sciences, once described as “extinct volcanoes”.

Summers is careful to avoid saying that he is seeking younger tenured faculty, but that appears to be at least a by-product of his philosophy — ‘potential’ sounds suspiciously like a euphemism for ‘youth’.

That's not to say that the age of tenured faculty — at Harvard or elsewhere — isn't an issue. At 47, Summers is eight years younger than the average tenured Harvard arts and science professor, and about two years younger than the US national average.

Of course, an increasingly ageing faculty isn't a problem in itself. But it can cause difficulties when tenured but unproductive professors cause obstructions for younger scholars seeking academic posts. I've heard many deans and department heads around the world wish — off the record, of course — that they had a way to remove an unproductive senior faculty member.

Summers deserves credit for making waves that could prompt other universities to consider hiring promising, if less experienced, faculty. But in the end, the decision to go for promise over the proven is a judgement call with an uncertain outcome. Even youthful volcanoes can become extinct. And dormant ones, if the conditions are right, can become active again.

Paul Smaglik

Naturejobs editor



Contents

CAREERS AND RECRUITMENT

The changing landscape in US ecology	p4
Europe tackles the Kyoto Protocol	p6
Building bridges in Japan	p7

WWW.NATUREJOBS.COM

Career centre
Information on the scientific job market

FOCUS

SPOTLIGHT

RECRUITMENT

SCIENTIFIC ANNOUNCEMENTS

SCIENTIFIC EVENTS



Fields of dreams

A new awareness of environmental problems is changing the landscape for scientists. In the United States, it is creating jobs for a diverse range of specialists in some unexpected areas. Potter Wickware reports.

“Ecology is rather like sex,” according to science writer Michael Allaby. “Every new generation likes to think they were the first to discover it.” Indeed, there is currently a sense of wide-eyed excitement surrounding the field as the latest generation gets to grips with its ‘discovery’. And this time around, solutions to the environment’s complicated problems are flowing from new combinations of scientific and professional talent.

Today’s wildlife biology, forestry or natural-resource group is as likely to include non-traditional posts such as environmental lawyers or mathematical modellers as more traditional roles such as geologists and telemetry specialists (see ‘A non-traditional approach’, right). And positions now tend to draw on several such backgrounds.

The term ‘ecological engineer’, for example, would describe someone who plans and develops ecologically sensitive sites, “someone grounded in biological systems who also has a good sense of how things work”, says Jason Taylor, education programme manager at the Ecological Society of America, a Washington-based professional society.

The social sciences increasingly enter the ecological equation too, adds Thomas Rice, a soil scientist at California Polytechnic University in San Luis Obispo. Rice recalls a student who came to him with an undergraduate degree in sociology, then turned to modelling the environmental effects brought about by mega-cities.

An increasing emphasis on building models for more environmental processes is causing some shortages: the

need for mathematicians and computer scientists who can aid in modelling is further exacerbated because they are also needed in other disciplines, such as computational biology. And climate science is often not the first choice or the best paid.

At the other end of the spectrum, ecological studies today depend more than ever on tools and techniques. Although solo investigators still count snails and lizards by hand at study sites, a lot of data now come from satellite imaging, aerial photography, computer simulations or ground-penetrating radar. Genomics offers entirely new data sets to help assess changes in communities as a whole, and for phylogenetic and coevolution studies, such as symbiotic relationships between birds and plants.

DRIVEN BY COMPLEXITY

It is no accident that more disciplines are intersecting at the ecological crossroads. That was the idea behind the biocomplexity initiative, an ambitious five-year programme set up by the National Science Foundation (NSF). It aims to foster core groups of scientists as they try to come to grips with ecological problems that range from the micrometre to the planetary scale.

The initiative’s conceptual origins were in physics and chaos theory. Then complexity was given a biological dimension by the work of NSF director Rita Colwell, a marine microbiologist, which associated cholera outbreaks with El Niño-driven changes in tidal-water temperatures.

The NSF currently supports biocomplexity research in five areas: coupled natural and human systems; coupled biogeochemical cycles; genome-enabled environmental science and engineering; instrumentation development for environmental activities; and materials science, engineering and society. In the current cycle 72 recipients share grants totalling about \$40 million.

In addition to independent jobs funded by research

Ecology now offers strong career prospects for a wide range of disciplines.



grants, there are many government staff positions in agencies such as the Agricultural Research Service (ARS), a branch of the US Department of Agriculture. The ARS has 8,000 staff scientists and support personnel dispersed among its 104 locations nationwide.

Mike Jawson, who leads programmes in natural resources and sustainable agriculture at ARS headquarters in Beltsville, Maryland, explains that in tandem with the traditional activities of stock and crop improvement, the ARS also works on water, soil and air quality and rangeland management. "Right now agro-ecologists — ecology-oriented systems people — are one of the main classifications we're looking to fill," he says. Highlighting ecology's hybrid complexion, recruitment adverts might have numerous headings for the same job: engineer, soil scientist, agronomist, hydrologist or ecologist.

ECOLOGICAL CHECKPOINTS

Weed ecologists and insect-pest ecologists occupy important positions under the agro-ecology umbrella. These scientists look at life cycles and behaviours and try to find checkpoints that can be exploited by prospective control agents.

"We have to make sure that the agent we're introducing doesn't harm something beneficial, and this takes a lot of research into the ecology of a region before a release can happen," says Jawson.

ARS ecologists are also active overseas. A collaborative effort in Brazil, for example, is developing its agriculture intensively while trying to avoid the mistakes made in the past by others. Erosion control in China and land rehabilitation in the former Soviet Republics of Central Asia are other big efforts.

Rangeland management is another area with bright jobs prospects. Some 40% of the United States is rangeland, and a typical problem for rangeland ecologists is what to do about the invasive salt cedar, which sucks streams dry but provides a home for a

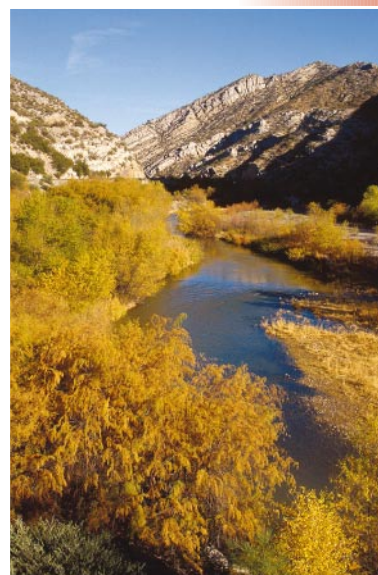
threatened bird, the southwestern willow flycatcher.

Other government agencies with staff ecologists are the Environmental Protection Agency, the US Forest Service, the Fish & Wildlife Service, and the Natural Resources Conservation Service (NRCS). The NRCS uses a lot of master's level people, in addition to a few PhDs, says Jawson. Even NASA has a finger in the ecology pie: since 1993 the Ames Research Center in Mountain View, California, has been involved in projects that apply multispectral image analysis to remote-sensing data to model the vigour of the vegetation canopy in vineyards.

Because agency field sites are apt to be situated on or near university campuses, government research jobs tend to have an academic cast. Joint agency and university appointments are common. Staff scientists enjoy an edge over their university counterparts in their guaranteed support base, which at the ARS averages \$300,000 per year per scientist. With less time spent chasing grant dollars, more time can be spent on research and field work.

The future is bright at the ARS for an ecologist who has taken a few agricultural courses and is conversant with the issues, says Jawson. Biotechnology skills are increasingly important, with proteomics slated to be a hot new part of research at the agency. "We have had to do double advertising to fill our positions more than once, particularly for these ecology-based systems positions," he says. "There's not a huge pool out there."

Potter Wickware is a freelance science writer based in California.



Delicate balance: the salt cedar drains streams but provides a home for endangered birds.

Scott Sellwood (below left) sees full employment for environmental lawyers.

A non-traditional approach

Ecologists' skills are increasingly in demand at environmental engineering companies. The consulting firm CH2M Hill, based in Denver, employs 400 scientists, most of them trained in Earth sciences, who work at the interface of agricultural and

natural systems, according to spokeswoman Karen Steeper. And traditional large construction firms, such as Bechtel of San Francisco, typically have environmental-compliance groups closely involved in

all stages of a project.

For the tech-minded, infrastructure companies that provide maps and data, and the computer methods to interpret them, also offer opportunities. Vestra, a land-management concern in Redding, California, employs foresters, geographers, cartographers and specialists in biometrics and remote sensing. ESRI, a geographic information consultancy, has 2,500 employees in Redlands, California, and other locations.

Thirty years ago, if you wanted to build in a wetland you began by calling in the gravel trucks. Today, projects in sensitive areas usually need

Environmental Impact Reports, and in many US states only a certified ecologist can sign these pivotal documents.

Damage caused by massive hazardous-waste dumps, now being cleaned up under Superfund legislation, will keep associated professions busy for years. According to Scott Sellwood, an environmental lawyer in San Francisco, related legal actions are giving lawyers a lot of work.

There is more litigation locked up in the Habitat Conservation Plans (HCPs), a comprehensive science-based method for managing endangered species. "HCPs are

a full-employment plan for consultants and lawyers," comments Sellwood.

Another good starting place for newcomers to ecology is an internship with a public-interest group or non-governmental organization. According to Jeffery Foran, president of Citizens for a Better Environment, a Great Lakes-based group, the experience can give a student a broader and more realistic perspective than is possible from behind university walls. "Advocacy is acceptable when done correctly and based on good strong scientific information," he judges.

P.W.



A climate of uncertainty



As ecologists attempt to get to grips with the Kyoto Protocol, many European climate scientists are concerned that there are insufficient resources available to understand the science behind the environment. Paul Smaglik reports.



Some ecologists worry that environmental research may have to be scaled back.

Ian Woodward (right) believes Britain is taking a balanced approach to climate-change research.



Although the Kyoto Protocol on Climate Change is raising the profile of ecologists, some are concerned that the work is emphasizing applied science at the expense of basic research. The European Commission seems to be supporting research that looks at the effects of global warming, while leaving member countries and others to investigate the mechanisms

behind the phenomenon, say climate scientists around Europe. Whether individual countries will fill that role may depend more on political and economic concerns than scientific ones.

Denmark, for one, may not be up to the task. The recently elected Conservative government has made signs that it will reduce government-funded environmental research as part of a scheme to provide some tax relief, says Tom Fenchel, director of Copenhagen University's Marine Biological Laboratory in Helsingør.

He worries that other European countries may follow suit. It is easier to be green in times of prosperity than in uncertainty, he says. Such signals may keep young scientists from entering a field that is already discouraging, as there are few positions available on university staff.

Over in England, his concerns are shared by Myles Allen of the Space Science Department at Rutherford Appleton Laboratory, Didcot. He thinks shifting too quickly to applied research could leave some of the questions on which it is based unanswered. "It's terribly damaging to assume that the

science is done just because you've signed the Kyoto Protocol."

Too much research on the impact of climate change, rather than on factors and dynamics, could discourage young mathematicians from entering ecology. The field could use those skills, because better quantitative measures could help to justify the case for taking action, and this would make any applied research that much more useful. Allen finds it ironic that the United States, which hasn't signed Kyoto, is more supportive of basic climate-change research than many of the countries that have agreed to the emission-reduction plan.

Ian Woodward, professor of the department of animal and plant sciences at Sheffield University, notes that Britain seems to be seeking a balanced approach. His lab, which examines the ecological impact of climate change on plant life and the role of plants in carbon sequestration, is one of five to share a five-year, £2.5-million (US\$7.2 million) Natural Environment Research Council (NERC) grant awarded this month to establish a Centre for Terrestrial Carbon Dynamics.

The centre will fund 10 additional postdocs at the participating institutions — a welcome addition, but hardly an overwhelming increase in the workforce. The award is guaranteed for five years — a long time in biology, but a mere blip in ecology.

One sign of commitment to applied research is the Tyndall Centre for Climate Change Research, at the University of East Anglia, Norwich. When the centre was established a year and a half ago, the Engineering and Physical Sciences Research Council, the NERC and the Economic and Social Research Council pledged an initial five-year grant of £10 million.

More long-term support for both basic and applied projects would help lend stability as well as balance, he says.

Paul Smaglik is editor of *Naturejobs*.

NERC ♦ www.nerc.ac.uk/funding/atmossoci

Tyndall Centre for Climate Change Research ♦ www.tyndall.ac.uk

MARINE BIOLOGICAL LAB, UNIV. COPENHAGEN

Interaction required

It may have profited from the increased interest in environmental issues, but ecology in Japan is still struggling to make ends meet. Robert Triendl crosses the divide within the discipline.

According to ecologists in Japan, two changes would improve their working lives. They would like to see more dialogue between basic and applied research, and they want to take a greater part in public debates and political decisions related to the environment. Like their colleagues in other industrialized countries, they have profited from the funding for ecological research that has increased along with concerns over issues such as biodiversity and global warming.

One example is Kyoto University's Centre for Ecological Research (CER), an inter-university research centre which is the first to be dedicated to ecology. Set up in 1992, the centre now hosts about 15 full and associate lecturers and some 70 postdoc and graduate students. Over the past few years, basic research in ecology has also benefited from a series of large-scale cooperative projects on ecosystems in Asia, such as tropical rainforests in Malaysia or Thailand. But most basic research in ecology in Japan remains confined to small university laboratories.

Research in ecology and environmental sciences was traditionally funded mainly by the Ministry of Education and, to a lesser extent, the former environmental agency. But during the early 1990s, agencies such as the New Energy Development Organization started to fund large-scale research efforts, often targeting applied research topics or risk assessment issues.

At the Institute for Environmental Sciences (IES) in Rokkasho Mura, where Japan's nuclear reprocessing plant is currently being built, scientists are building a small artificial habitat, similar to the US Biosphere 2 project. With an annual budget of some US\$50 million, the IES is now one of the best-funded environment-related research centres in Japan. After completion, scientists will trace the accumulation of radioactive substances in the environment.

But as Keiji Nitta, who directs the environmental simulation department at the IES, points out, the institute's facilities could be used to study a host of



other questions. Given the institute's background and its source of funding, research on topics outside its immediate mission and collaborations with external scientists have not been encouraged. This is now slowly changing, says Nitta.

But collaboration between academic scientists and researchers working on more applied questions is not straightforward. Ecologists in Japan need to broaden their view and get more involved in questions relevant to real-world environmental issues, argues Nitta. Ecologists in Japan have tended to be highly specialized, with only limited interaction among the various branches of the field.

Norio Yamamura, director of the CER, agrees that many ecologists in Japan have been reluctant to engage in research on high-profile issues or to connect with policy-makers.

For young scientists there is often little choice. As an academic discipline, the job prospects for aspiring ecologists are not encouraging. The number of faculty positions is hardly growing and the number of research positions for ecologists outside academia is limited. The situation looks much better outside the realm of purely academic research.

The challenge remains to bridge these two worlds. More interaction between basic and applied research and higher visibility by the general public and political decision-makers will eventually benefit the whole field.

Robert Triendl is a freelance writer based in Tokyo.

Japanese ecologists would like to build more bridges between the basic and applied parts of the discipline.